2020

Algorithms and Applications of Novel Capsule Networks

Rodney LaLonde
University of Central Florida

ALGORITHMS AND APPLICATIONS OF NOVEL CAPSULE NETWORKS

by

RODNEY LALONDE III
M.Sc. University of Central Florida, 2019
B.A. St. Olaf College, 2016

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Summer Term
2020

Major Professor: Ulas Bagci

ABSTRACT

Convolutional neural networks, despite their profound impact in countless domains, suffer from significant shortcomings. Linearly-combined scalar feature representations and max pooling operations lead to spatial ambiguities and a lack of robustness to pose variations. Capsule networks can potentially alleviate these issues by storing and routing the pose information of extracted features through their architectures, seeking agreement between the lower-level predictions of higher-level poses at each layer.

In this dissertation, we make several key contributions to advance the algorithms of capsule networks in segmentation and classification applications. We create the first ever capsule-based segmentation network in the literature, ***SegCaps***, by introducing a novel locally-constrained dynamic routing algorithm, transformation matrix sharing, the concept of a "deconvolutional" capsule, extension of the reconstruction regularization to segmentation, and a new encoder-decoder capsule architecture. Following this, we design a capsule-based diagnosis network, ***D-Caps***, which builds off *SegCaps* and introduces a novel capsule-average pooling technique to handle to larger medical imaging data. Finally, we design an explainable capsule network, ***X-Caps***, which encodes high-level visual object attributes within its capsules by utilizing a multi-task framework and a novel routing sigmoid function which independently routes information from child capsules to parents. Predictions come with human-level explanations, via object attributes, and a confidence score, by training our network directly on the distribution of expert labels, modeling inter-observer agreement and punishing over/under confidence during training. This body of work constitutes significant algorithmic advances to the application of capsule networks, especially in real-world biomedical imaging data.

EXTENDED ABSTRACT

Deep learning methodologies, in particular convolutional neural networks (CNNs), have made a profound impact in countless domains across academia, government, and industry. Nonetheless, over the past few decades, some have argued strongly against their core mechanisms. Linearly-combined scalar feature representations and max pooling operations lead to spatial ambiguities and a lack of robustness to pose variations. Capsule networks are a class of neural networks which aim to solve these shortcomings by storing both presence and pose information about extracted features, and route this information through the network seeking agreement between lower- and higher-level features. In this dissertation, we make several key contributions to advance the algorithms and application of capsule neural networks. Specific focus is given to biomedical image applications for their significance in potentially life-saving technologies.

Discussed in Chapter 3, we introduce the first ever capsule network designed for the task of segmentation in the literature. This required several important advancements, including a novel locally-constrained dynamic routing algorithm, transformation matrix sharing, the concept of a "deconvolutional" capsule, extension of the reconstruction regularization to segmentation, and a new encoder-decoder capsule network structure. These advancements culminate in an architecture which we call ***SegCaps***. *SegCaps* consistently outperforms contemporary state-of-the-art CNNs in pathological lung segmentation for both clinical and preclinical subjects, as well as retinal vessel segmentation from fluorescein angiogram videos, while using only a small fraction of the trainable parameters as those CNNs. Further, we provide evidence that *SegCaps* can generalize to unseen poses of familiar objects far better than a state-of-the-art CNN.

Next, in Chapter 4, we design a capsule-based network for the task of diagnosis in the field of endoscopy. In order to classify real-world imaging data much larger in size than those in MNIST or

CIFAR, we introduced the concept of *capsule-average pooling*. Our proposed architecture, which we call ***D-Caps***, combines this capsule-average pooling with the parameter saving techniques introduced in *SegCaps* to diagnose colorectal polyps from colonoscopy images. Given our preliminary evidence that capsule networks can better generalize to unseen poses, converges faster in training, and contains far fewer parameters than state-of-the-art CNNs, we hypothesize that *D-Caps* should be able to better handle the relatively limited training data and high intra-class variation present in our colorectal polyp dataset. We conducted a set of thorough experiments to validate our hypothesis, stratified across all polyp categories, imaging devices and modalities, and focus modes available. Our results show *D-Caps* can outperform the leading state-of-the-art CNN-based method by as much as 43% in the most difficult settings.

In Chapter 5, we introduce algorithmic advances in capsule networks to improve the explainability of network predictions. CNN-based systems have largely not been adopted in many high-risk application areas, including healthcare, military, security, transportation, finance, and legal, due to their highly uninterpretable "black-box" nature. Towards solving this deficiency, we teach a capsule network to explain its predictions using the same high-level language used by human-experts. Our explainable capsule network, ***X-Caps***, encodes high-level visual object attributes within the vectors of its capsules, then forms predictions based solely on these human-interpretable features. We implement a multi-task learning framework to learn the attribute and malignancy scores from a large multi-center dataset of lung cancer screening patients. *X-Caps* utilizes a routing sigmoid to independently route information from child capsules to parents for the visual attribute vectors. To estimate model confidence, we train our network on a distribution of expert labels, modeling inter-observer agreement and punishing over/under confidence during training supervised by human-experts' agreement. We demonstrate a simple 2D capsule network trained from scratch can outperform a state-of-the-art deep pre-trained dense dual-path 3D CNN at capturing visually-interpretable high-level attributes and malignancy prediction, while providing malignancy prediction

scores approaching that of non-explainable 3D CNNs.

The body of work presented in this manuscript constitutes significant algorithmic advances to the application of capsule networks in a variety of real-world imaging data domains, and in particular, biomedical image computer-aided diagnosis. Capsule networks show considerable promise for the future of deep learning-based applications and we hope the contributions of this dissertation provide a solid foundation for the further advancement of capsule network-based approaches. The source code for all of the algorithms discussed have been made publicly available at https://github.com/lalonderodney.

For my mother, father, and brother who always supported me. You sacrificed so much to give me a chance at a better life. This is the long culmination of a dream for me that we all shared.
I know this document could never have been achieved without all that you have done for me, and so I dedicate this work to you. I love you each of you so much.

ACKNOWLEDGMENTS

I would like to give my deepest appreciation to my Ph.D. adviser Dr. Ulas Bagci for all of the teaching, advice, and training over the last several years. You taught me how to be a true scientist and believed in me since you first read my application letter to UCF.

I would like to thank Dr. Mubarak Shah for co-advising me during the first year of my Ph.D. and the great experience he instilled into me. Also, to Dr. Dong Zhang, thank you for the guidance during my first year.

To the other members of my committee, Dr. Abhijit Mahalanobis and Dr. Michael B. Wallace, I would like to thank you for your commitment, guidance, and support throughout my Ph.D. process.

To all of my co-authors over the years, I would like to thank you for the thoughtful help you gave in shaping our work and the time and dedication you showed in providing crucial data and annotations on display in this manuscript. Specifically I would like to give my thanks to Dr. Michael B. Wallace, Dr. Pujan Kandel, and Dr. Candice W. Bolan from the Mayo Clinic, Jacksonville, to Dr. Drew Torigian from the University of Pennsylvania, to Dr. Concetto Spampinato from the University of Catania, and to Dr. Sanjay Jain from Johns Hopkins University.

To my lab mates, and specifically those in my group with Dr. Bagci: Dr. Sarfaraz Hussein, Dr. Naji Khosravan, Dr. Ali Mortazi, and Harish RaviPrakesh, I would like to thank you all for keeping me sane and for all the great discussions over the years on many research ideas.

I would like to give a special thank you to the funding agencies that made my Ph.D. possible, including Lockheed Martin, the Florida Department of Health, the National Institute of Health subcontracts, and the University of Central Florida for the ORC Doctoral Fellowship.

Lastly, and most importantly, I would like to give thanks to all those who loved and supported me over the last several years pursuing this degree. To my fiancé Catherine O'Leary, thank you for the endless support you showed when I spent countless sleepless nights in the lab pursuing these works. To my parents and brother, thank you for the frequent calls filled with supportive reassurances and for helping to keep me fed on my shoestring grad student budget. To all those friends from my past not named in this document, know that you are not forgotten, and though you may not have directly helped during my Ph.D., I would have never made it here without every one of you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Deep learning methodologies have made a profound impact in countless domains across academia, government, and industry. The great success of these methods in part can be attributed to the hierarchical representations of features extracted by the many layers of non-linear functions stacked to form a neural network, hence the term deep learning. As a general rule, the deeper, the better, to within some limits [47]. To create deeper and more powerful networks, fully-connected multi-layer perceptrons were replaced by convolutional neural networks (CNNs), which reuse kernels across spatial dimensions and store extracted features within scalar feature maps. To reduce computational burdens further, max-pooling layers were added which reduce the spatial dimensions of feature maps by extracting only the maximum values within local neighborhoods. Every year, deeper and more powerful CNNs were introduced [67, 111, 115]. Nonetheless, over the past few decades, the "godfather of deep learning" Geoffrey Hinton [112, 120], has argued strongly against some of their core mechanisms, favoring instead what he calls *capsule networks* [48].

1.1 Drawbacks of CNNs and How Capsules Solve Them

Convolutional neural networks, despite showing remarkable flexibility and performance in a wide range of computer vision tasks (*e.g.* classification [53], detection [101], segmentation [79]), do come with their own set of flaws. Due to the scalar and additive nature of neurons in CNNs, neurons at any given layer of a network are ambivalent to the spatial relationships of neurons within their kernel of the previous layer, and thus within their effective receptive field of the given input. This is worsened by the introduction of max-pooling which further destroys the spatial relationships between features. To address these significant shortcomings, capsule networks store information at the neuron level as vectors, rather than scalars. These vectors contain information about the

extracted features, including prevalence, pose, color, scale, and more, represented by each dimension of the capsule vector. In this way. capsules provide *equivariance* to affine transformations on the input, as opposed to CNNs which are only equivarient to translation. These sets of neurons, are then "routed" to capsules in the next layer via a *dynamic routing algorithm* which takes into account the agreement between these capsule vectors, thus forming meaningful part-whole relationships not found in standard CNNs.

A simple three-layer capsule network, called *CapsNet* [105], showed remarkable initial results, producing state-of-the-art classification results on the MNIST dataset and relatively good classification results on the CIFAR10 dataset. Since then, researchers have begun extending the idea of capsule networks to other applications, including brain-tumor classification [2], lung-nodule screening [85], action detection [24], point-cloud autoencoders [129], adversarial detection [38, 96], and even creating wardrobes [52], as well as several technical contributes to improve the routing mechanism for datasets such as MNIST, CIFAR10, SVHN, SmallNorb, etc. [49, 66].

1.2 Segmenting Objects from Images with Capsules

Object segmentation in the medical imaging and computer vision communities has remained an interesting and challenging problem over the past several decades. The task of segmenting objects from images can be formulated as a joint object recognition and delineation problem. The goal in *recognition* is to locate an object's presence in an image, whereas delineation attempts to draw the object's spatial extent and composition [8]. Solving these tasks jointly (or sequentially) results in partitions of non-overlapping, connected regions, homogeneous with respect to some signal characteristics. Object segmentation is an inherently difficult task; apart from recognizing the object, we also have to label that object at the pixel level, which is an ill-posed problem.

In Chapter 3, the overall goal is to extend the concept of capsule networks to accomplish the task of object segmentation for the first time in the literature. We hypothesize that capsules can be used effectively for object segmentation with high accuracy and heightened efficiency compared to the state-of-the-art CNN-based segmentation methods. This required several important advancements, including a novel locally-constrained dynamic routing algorithm, transformation matrix sharing, the concept of a "deconvolutional" capsule, extension of the reconstruction regularization to segmentation, and a new encoder-decoder capsule network structure. These advancements culminate in an architecture which we call *SegCaps*. In short, locally-constrained dynamic routing constrains parent capsules to only receive information from a small local neighborhood of child capsules centered on the parent's position. This dramatically reduces the size of the transformation matrices, and to reduce the memory burden further, we share transformation matrices across each member of the grid. To compensate for the loss of global information, we adopt an encoder-decoder network structure, which is facilitated by the introduction of deconvolutional capsules. These deconvolutional capsules are similar to convolutional capsules, except their prediction vectors are formed via a transposed convolution operation. Lastly, we extend the reconstruction regularization, shown to be effective in capsule networks, by reconstructing a class-wise masked version of the input.

To demonstrate the efficacy of *SegCaps*, we choose a challenging application of pathological lung segmentation from computed tomography (CT) scans, where we have analyzed the largest-scale study of data obtained from both clinical and pre-clinical subjects, comprising nearly 2000 CT scans across five datasets. While our methods presented are applied to biomedical data, we want to emphasize that our method is in no way specific to medical imaging. We chose pathological lung segmentation for its obvious life-saving potential and unique challenges such as high intra-class variation, noise, artifacts and abnormalities. To further demonstrate the general applicability of our methods, we also provide proof-of-concept results in retinal vessel segmentation from fluorescein angiogram videos which contain extremely thin tree-like structures as well as for

rotations/reflections on standard computer vision images. These experiments provide evidence that *SegCaps* can generalize to unseen poses of familiar objects far better than a state-of-the-art CNN. *SegCaps* consistently outperforms contemporary state-of-the-art CNNs, while using only a small fraction of the trainable parameters as those CNNs, showing a strong motivation for choosing capsule networks over CNNs in segmentation applications.

1.3 Capsules for Computer-Aided Diagnosis

The memory saving methodologies introduced in *SegCaps* allowed for segmentation at real world image sizes (512×512 pixels, as compared to 28×28 in previous studies). However, the fully-convolutional structure of *SegCaps* does not easily allow for classification tasks at the same scales. To overcome this technical shortcoming, in Chapter 4 we introduce the concept of a *capsule-average pooling* (CAP) function. In theory, CAP behaves in a similar manner to the global average pooling function in CNNs and acts to reduce the spatial dimensions of capsule layers. More explicitly, this novel algorithm computes the average value along individual capsule dimensions, across the capsule grid, separately for each capsule type. In this way, we compute a single capsule vector to represent the entity being modeled by each capsule type for each layer the CAP function is applied to, regardless of spatial position. From these single entity vectors, we can then efficiently perform classification on large scale images.

Our proposed architecture, which we call *D-Caps*, combines the parameter saving techniques introduced in *SegCaps* with our new CAP algorithm to diagnose colorectal polyps from colonoscopy images. This application is chosen for several reasons: 1) colorectal cancer is a leading cause of cancer-related death, 2) colorectal polyp datasets are relatively small, containing only a few hundred examples, and 3) the variation in location, scale, shape, illumination, and color of polyps makes the task extremely challenging. Given our preliminary evidence that capsule networks can better

generalize to unseen poses, converges faster in training, and contains far fewer parameters than state-of-the-art CNNs, we hypothesize that *D-Caps* should be able to better handle the relatively limited training data and high intra-class variation present in our colorectal polyp dataset.

To validate our hypothesis, we conducted a set of thorough experiments on the Mayo Polyp dataset [28]. As opposed to more unrealistic "academic" datasets, this dataset is far closer to true clinical settings, with only a single image per imaging mode per polyp, large inter-polyp variation, and often only a single imaging mode provided. From these experiments, we provide results stratified across all polyp categories, imaging devices and modalities, and focus modes available. Our analysis shows *D-Caps* can outperform the leading state-of-the-art method [16], based on Inceptionv3 [116], by as much as 43% in the most difficult settings, while using 95% fewer trainable parameters.

1.4 Creating an Explainable Capsule Network

Although the creation of a more effective and efficient diagnosis network, *D-Caps*, is surely an important step toward computer-aided diagnosis (CAD) systems being adopted into routine clinical workflows, there is a more significant barrier that also must be overcome. Recent CNN-based CAD systems have obtained remarkable performance, even exceeding human-experts in certain applications; nonetheless, they are largely not adopted into clinical workflows. This same hesitancy is seen in many high-risk application areas, including military, security, transportation, finance, and legal. The most cited reason behind this reluctance of adoption is lack of trust, caused by the highly uninterpretable "black-box" nature of CNNs. In Chapter 5, we introduce algorithmic advances in capsule networks to improve the explainability of network predictions.

Towards solving this deficiency, we teach a capsule network to explain its predictions using the same high-level language used by human-experts. As an example application, we look at lung

cancer diagnosis from computed tomography scans. In this domain, the high-level language used by human-experts is six high-level visual attributes, scored by radiologists on a scale from 1 to 5. These attribute scores provide the basis by which radiologists determine their final diagnoses. Our explainable capsule network, which we refer to as *X-Caps*, encodes these high-level visual object attributes within the vectors of its capsules, then forms predictions based solely on these human-interpretable features.

We implement a multi-task learning framework to simultaneously learn the attribute and malignancy scores for a large multi-center dataset of lung cancer screening patients. Since our capsule types are no longer mutually exclusive, we need to modify the dynamic routing algorithms to support this new formulation. *X-Caps* utilizes a novel modification to the dynamic routing algorithm based on a routing sigmoid principle. This enables child capsules to independently route information to parents for each of the visual attribute vectors. To further increase the explainability of our method, we propose to train our network directly on the distribution of expert labels, modeling inter-observer agreement, rather than their average as done in previous studies. At test, this provides a meaningful metric of model confidence, punishing over/under confidence during training supervised by human-experts' agreement, while visual attribute prediction scores are verified via a reconstruction branch of the network.

We validated our proposed method with experiments conducted on a large scale multi-center dataset of lung cancer screening patients. Our proposed *X-Caps* demonstrates that a simple 2D capsule network trained from scratch can outperform a state-of-the-art deep pre-trained dense dual-path 3D CNN at capturing visually-interpretable high-level attributes and malignancy prediction, while providing malignancy prediction scores approaching that of non-explainable 3D CNNs.

CHAPTER 2: LITERATURE REVIEW

The following three sections correspond to the work detailed in Chapters 3– 5 respectively. Section 2.1 describes the works in image segmentation from pre-deep learning to current state-of-the-art approaches. Section 2.2 covers the most recent advances in capsule neural networks as applied to the task of medical image diagnosis, as well as all works related to automated colorectal polyp diagnosis. Finally, Section 2.3 describes a large body of literature in the areas of explainable deep learning, lung nodule classification, and their intersection. Since explainable deep learning is a relatively new field of study, there is a fairly diverse range of approaches which have been proposed, and we try to faithfully cover the most significant works in each of these thrusts, finishing with the proposed methods in explainable lung cancer diagnosis.

2.1 Image Segmentation

Early attempts in automated object segmentation were analogous to the if-then-else expert systems of that period, where the compound and sequential application of low-level pixel processing and mathematical models were used to build-up complex rule-based systems of analysis [50, 103]. In computer vision fields, superpixels and various sets of feature extractors such as scale-invariant feature transform (SIFT) [81] or histogram of oriented gradients (HOG) [20] were used to construct these spaces. Specifically in medical imaging, methods such as level sets [121], fuzzy connectedness [119], graph-based [35], random walk [44], and atlas-based algorithms [94] have been utilized in different application settings. Over time, the community came to favor supervised machine learning techniques, where algorithms were developed using training data to teach systems the optimal decision boundaries in a constructed high-dimensional feature space.

In the last few years, deep learning methods, in particular convolutional neural networks (CNNs), have become the state-of-the-art. Specifically related to the object segmentation problem, *U-Net* [102], Fully Convolutional Networks (FCN) [79], and other encoder-decoder style CNNs (*e.g.* [87]) have become the desired models for various medical image segmentation tasks. Most recent attempts in the computer vision and medical imaging literature utilize the extension of these methods to address the segmentation problem [128, 15, 124]. Herein, we only summarize the most popular deep learning-based segmentation algorithms.

Based on FCN [79] for semantic segmentation, *U-Net* [102] introduced an alternative CNN-based pixel label prediction algorithm which forms the backbone of many deep learning-based segmentation methods in medical imaging today. Following this, many subsequent works follow this encoder-decoder structure, experimenting with dense connections, skip connections, residual blocks, and other types of architectural additions to improve segmentation accuracy for particular imaging applications. For instance, a recent example by Jégou et al. [58] combines a *U-Net*-like structure with the very successful DenseNet [53] architecture, creating a densely connected *U-Net* structure, called *Tiramisu*. Another example, Mortazi et al. [88] proposed a multi-view CNN, following this encoder-decoder structure and adding a novel loss function, for segmenting the left atrium and proximal pulmonary veins from MRI.

Among the most recent successes, SegNet [7] attempts to improve the upsampling process by performing "unpooling", capturing the pooling indices from the max pooling layers in the encoder to more accurately place features in the decoder feature maps. Although the encoder-decoder structure is specifically designed to capture global context information, several methods attempt to further improve this global context in different ways. RefineNet [78] fuses features from multiple resolutions through adding residual connections and chained residual pooling to create a large cascaded encoder-decoder structure. PSPNet [128] introduces a pyramid pooling module by pooling at different kernel sizes and concatenating back to the features maps. Large Kernel Matters [93] uses

large $1 \times 15 + 15 \times 1$ and $15 \times 1 + 1 \times 15$ global convolution networks. ClusterNet [72] combines two fully-convolutional networks, one to capture global and one for local information, to segment specifically a large number of densely packed tiny objects, normally lost in networks with pooling. DeepLab [14] utilizes an atrous spatial pyramid pooling (ASPP) unit to better capture image context from multiple scales. The latest version of DeepLab (v3+) [15] follows a very similar structure to *U-Net* with the addition of an ASPP for image context and depthwise separable convolutions for efficiency. Specific to pathological lung segmentation, *P-HNN* [46], achieved very strong results on a subset of three clinical datasets by modifying the Holistically-Nested Network (HNN) [123] structure to progressively sum side-output predictions during the decoder phase.

2.2 Capsule-Based Methods & Colorectal Polyp Diagnosis

Advances in capsule networks: Since the initial publication by Sabour et al. [105], there has been an explosion of capsule-based research methods to appear in the literature. The majority of these methods have focused on image-based classification [21, 49, 97, 122], with some other notable work in areas such as action detection [24, 25, 84], point-cloud autoencoders [129], adversarial detection [38, 96], similarity matching/image retrieval [52, 65], generative methods [57], and reinforcement learning [5]. The bulk of these recent studies have been primarily application focused in their novelty; however, several studies have attempted to advance the algorithms of capsule networks, primarily focusing on the dynamic routing mechanism [49, 66, 91, 118].

Capsule networks for medical image diagnosis: A number of recent studies have proposed using *CapsNet* [105] for a variety of medical imaging classification tasks [2, 56, 86, 92, 110]. However, these methods nearly all follow the exact *CapsNet* architecture with their novelty lying solely in the application of the network to new datasets and domains. For example, Jiménez-Sánchez et al. [59] employed CapsNet for a number of medical and non-medical tasks, and show some early evidence

that capsule networks may generalize better given limited data. Of those works which do propose novel modifications to the *CapsNet* formula, they typically only propose very minor modifications which nonetheless present nearly identical predictive performance [85].

Background on automated colorectal polyp diagnosis: Specific to colorectal polyp diagnosis, the number of computer-aided diagnosis studies is somewhat limited. In [32], a bag-of-features representation was constructed by a hierarchical k-means clustering of scale invariant feature transform (SIFT) descriptors. These features were then used to train an SVM classifier for classifying hyperplastic polyps vs adenomas. The approach by [31] was the first to incorporate deep learning to diagnose hyperplastic polyps vs adenomas. The authors extracted the first 3 – 4 layers of an Inception-style network trained on ImageNet and Places205 and trained an SVM to classify the extracted deep features. The first end-to-end trained network was used in [29], which employed an AlexNet style network trained from scratch with data augmentation to classify polyps as hyperplastic, adenomas, none, or unsuitable image. Most recently, [16] used a pretrained Inceptionv3 network to classify hyperplastic polyps from adenomas.

2.3 Explainable Deep Learning for Medical Diagnosis

The majority of work in explainable deep learning has focused around *post hoc* deconstruction of already trained models. Two main approaches are primarily investigated, interpretation of the features learned by the networks and explaining deep networks' final predictions, at both the local (*i.e.* individual neurons) and global (*i.e.* entire layers/networks) levels. These approaches typically rely on human-experts to examine their results and attempt to discover meaningful patterns. While there are numerous studies on interpretable and explainable DL, we will attempt to faithfully cover the more prominent approaches. Following this, we will cover relevant lung cancer diagnosis and capsule-based works.

Visualization of features: Several works have attempted to examine network interpretability at the individual neuron level. Some of the earliest methods focused on visualizing individual filters and activation maps. While this can provide some insight into aspects of a network, such as dead neurons, the visualization of individual filters or feature maps are typically not interpretable at the human-level. Zeiler and Fergus [126] attached a deconvolutional network to network layers to map activations back to pixel space for visualization. Later, Springenberg et al. [113] used an all convolutional network and a guided-backpropagation algorithm to create much sharper visualizations which did not require the keys of the pooling operations. Mahendran and Vedaldi [83] focused more on layers of neurons and examine the representations learned by shallow and deep CNNs by inverting images using gradient descent. While these methods provide some insight into what CNNs learn, they are ultimately limited, as deep networks typically have hundreds of thousands of neurons and it is intractable to visually examine all or even large subsets of neurons in a network. Additionally, there is evidence to suggest these visualizations are unrelated to network predictions [90].

Receptive fields, input contributions: Beyond visualizing the features of CNNs, several methods have attempted to examine the effect of individual neurons or image regions on network outputs. In this first category, Girshick et al. [42] examined the receptive field of individual neurons and found the images which maximally activated each. Long et al. [80] showed that CNNs actually localize features at a much smaller scale than their theoretical receptive field. Zhou et al. [130] developed a method for visualizing these "empirical" receptive fields of neurons. Kindermans et al. [63] showed that Springenberg et al. [113] and Zeiler and Fergus [126] (discussed above) did not create theoretically correct explanations for linear models, and created *PatternNet* and *PatternAttribution* to better visualize neuron activations. In the latter category, Kumar et al. [69] examined which input region correspond most strongly with each output class. Similarly, Zintgraf et al. [132] examined which image regions contributed most negatively and positively to the correct classification score.

An occlusion-based approach was used by Zeiler and Fergus [126] for masking out image regions to examine their contribution to the final output. One of the most popular methods of visualizing input contributions is *Grad-CAM* [106] which highlights the relative positive activation map of convolutional layers with respect to network outputs. Arguably, saliency detection can also fall into this category of determining input region importance. While these methods give important information related to designing networks and training data, they tell us very little about the internal representations being learned.

Feature spaces and GANs: Rather than looking at the individual neurons or image regions, several approaches instead focus on examining the feature spaces learned by deep networks. Generative adversarial networks (GAN) by Goodfellow et al. [43], show vulnerable regions of a learned feature space for a given network. Chen et al. [17] creates a GAN-based method called *InfoGAN* to separate noise from the "latent code" in images. Using this method, they maximize the mutual information between the latent representations and the image inputs, encoding concepts such as rotation, width, and digit type for MNIST. In a similar way, capsule networks by Sabour et al. [105] (*CapsNet*) encode visually-interpretable concepts such as stroke thickness, skew, rotation, and others. These two methods are the most similar to the proposed approach. Lakkaraju et al. [70] attempt to discover a CNN's "blind spots" by sampling points in feature space in a weakly-supervised manner. From a purely visualization point of view, t-SNE [82] is often used to visualize the high dimensional feature spaces learned by deep networks. While the other methods mentioned can provide some important clues about the feature space being learned, *InfoGAN* and *CapsNet* show the most promise for encoding and extracting visually-interpretable features.

Disentangling representations: Methods for disentangling representations are focused on discovering the visual patterns learned by CNN filters, then disentangling their relationship to each other. Zhang et al. [127] created multi-layer graph structure, where each layer of the graph matches each layer of the CNN. Activated visual patterns across all training images are added as nodes and

patterns which co-occur in images have edges added between them. Bau et al. [9] introduced six types of semantic filters for CNNs: objects, parts, scenes, textures, materials, and colors. Networks are then trained using these labels at the pixel-level to identifies hidden units' semantics for any given CNN, and align them with human-interpretable concepts. Unfortunately, the former of these methods can only provide little about the features learned, while the latter method requires a dramatic increase in labeled data, where multiple labels need to be provided at the pixel level.

Lung nodule classification: The majority of recent lung cancer diagnosis (nodule classification) studies have focused on deep 2D, multi-view, and 3D CNNs, with most works trained/tested on the publicly available LIDC-IDRI data set from Lung Image Database Consortium [6]. Buty et al. [13] extracted features from a pre-trained 2D multi-view CNN while encoding shape information though spherical harmonics (SH) to improve diagnostic accuracy from 79% (CNN) to 82% (CNN+SH). Hussein et al. [55] achieved a similar result, extracting deep features from a multi-view CNN then applying a Gaussian process regression strategy to achieve 82% accuracy. Li et al. [77] used a 3D deep CNN MTL learning approach, where attributes were predicted along with malignancy, and achieved a diagnosis accuracy of $80-83\%$, depending on the visual attributes chosen, although again no results were reported on the accuracy of predicting attributes.

Explainable lung cancer diagnosis: More recently, some deeper multi-crop [109], multi-scale [108], and denser dual-path multi-output [23] 3D CNNs, using methods such as curriculum learning [89] or gradient boosting machines [131] and complicated post-processing techniques [54], have been applied to push diagnosis accuracy to $87\% - 92\%$. However, adding such techniques is beyond the scope of this work and would lead to an unwieldy enumeration of ablation studies necessary to understand the contributions between our proposed capsule architecture and such techniques. For a fair comparison in this study, we compare our method directly against *CapsNet* and explainable CNN approaches. Shen et al. [107] is one of the only works in the literature to attempt to create an interpretable framework by simultaneously predicting visual attribute scores along with malignancy.

This decreased the overall performance as compared to other 3D networks [109] but provided some explanations for the final malignancy predictions. The authors used a deep dual-path dense 3D CNN to achieve an accuracy of 84%, however their results on individual attribute predictions were as low as 55%.

CHAPTER 3: CAPSULES FOR IMAGE SEGMENTATION

Related Publications and Patents:

- Rodney LaLonde and Ulas Bagci. "Capsules for Object Segmentation." *MIDL 2018, Medical Imaging with Deep Learning*. 2018. (CIFAR Travel Award for Outstanding Papers). [71]
- Rodney LaLonde, Ziyue Xu, Sanjay Jain, and Ulas Bagci. "Capsules for Biomedical Image Segmentation." *Medical Image Analysis; Under Revision*. [75]
- Sumit Laha, Rodney LaLonde, Austin E. Carmack, Hassan Foroosh, John C. Olson, Saad Shaikh, and Ulas Bagci. "Analysis of Video Retinal Angiography with Deep learning and Eulerian Magnification." *Frontiers in Computer Science; In Press*.
- Rodney LaLonde and Ulas Bagci. "Capsules for Image Analysis." *U.S. Patent Application 16/431,387; filed December 5, 2019*.

In this chapter, we focus on the first major extension of *CapsNet* [105] by designing a deep encoder-decoder capsule network for the task of object segmentation for the first time in the literature. The remainder of this chapter is organized into the following sections: Section 3.1 – Building blocks which are invented to create our capsule-based segmentation framework, including the locally-constrained dynamic routing and transformation matrix sharing; Section 3.2 – The *SegCaps* framework described in detail, including the deconvolutional capsules and reconstruction regularization for segmentation; Section 3.3 – Experiments conducted, our implementation settings, and their results; Section 3.4 – Ablation studies on the novel components of our algorithms to determine the contribution of each aspect of our proposed method to the final results; Section 3.5

– Additional experiments with other types of imaging data and applications to provide empirical support for the general applicability of our study; and Section 3.6 – Discussions and concluding remarks.

3.1 Building Blocks of Capsules for Segmentation

Performing object segmentation with a capsule-based network is difficult for a number of reasons. The original capsule network architecture and dynamic routing algorithm is extremely computationally expensive, both in terms of memory and run-time. Additional intermediate representations are needed to store the output of "child" capsules in a given layer while the dynamic routing algorithm determines the coefficients by which these children are routed to the "parent" capsules in the next layer. This dynamic routing takes place between every parent and every possible child. One can think of the additional memory space required as a multiplicative increase of the batch size at a given layer by the number of capsule types at that layer. The number of parameters required quickly swells beyond control as well, even for trivially small inputs such as MNIST and CIFAR10. For example, given a set of 32 capsule types with 6×6, 8D-capsules per type being routed to 10×1, 16D-capsules (as is the case in CapsNet), the number of parameters for this layer alone is $10 \times (6 \times 6 \times 32) \times 16 \times 8 = 1,474,560$ parameters. This one layer contains, coincidentally, roughly the same number of parameters as our entire proposed deep convolutional-deconvolutional capsule network with locally-constrained dynamic routing which itself operates on 512×512 pixel inputs.

We solve this memory burden and parameter explosion by extending the idea of convolutional capsules (primary capsules in *CapsNet* [105] are technically convolutional capsules without any routing) and rewriting the dynamic routing algorithm in two key ways. First, children are only routed to parents within a defined spatially-local kernel. Second, transformation matrices are shared for

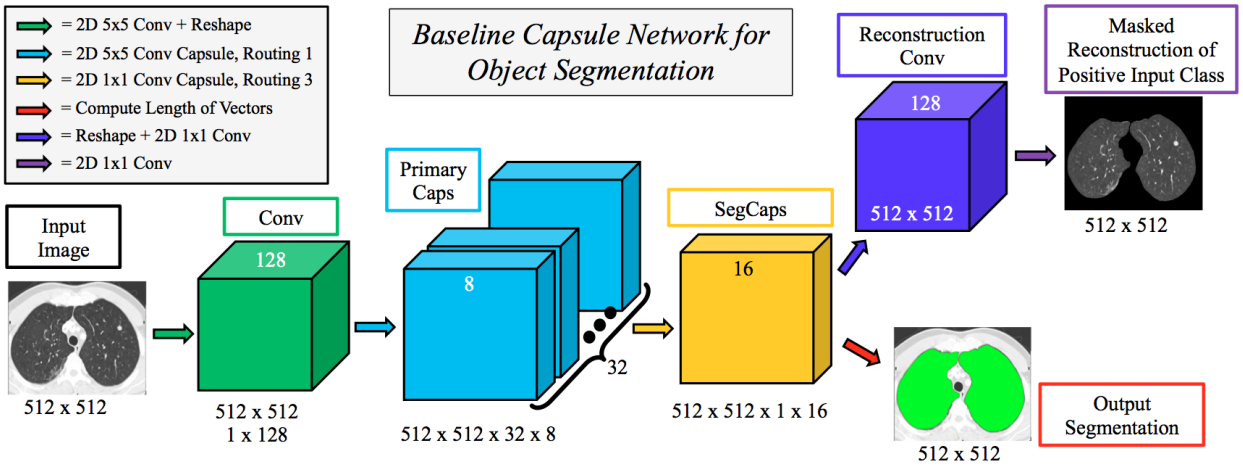


Figure 3.1: A simple three-layer capsule segmentation network closely mimicking the work by Sabour et al. [105]. This network uses our proposed locally-constrained dynamic routing algorithm as well as the masked reconstruction of the positive input class.

each member of the grid within a capsule type but are not shared across capsule types. To compensate for the loss of global connectivity with the locally-constrained routing, we extend capsule networks by proposing "deconvolutional" capsules which operates using transposed convolutions, routed by the proposed locally-constrained routing. These innovations allow us to still learn a diverse set of different capsule types while dramatically reducing the number of parameters in the network, addressing the memory burden. Also, with the proposed deep convolutional-deconvolutional architecture, we retain near-global contextual information and produce state-of-the-art results for our given application. Our proposed ***SegCaps*** architecture is illustrated in Figure 3.2. As a comparative baseline, we also implement a simple three-layer capsule structure, more closely following that of the original capsule implementation, shown in Figure 3.1.

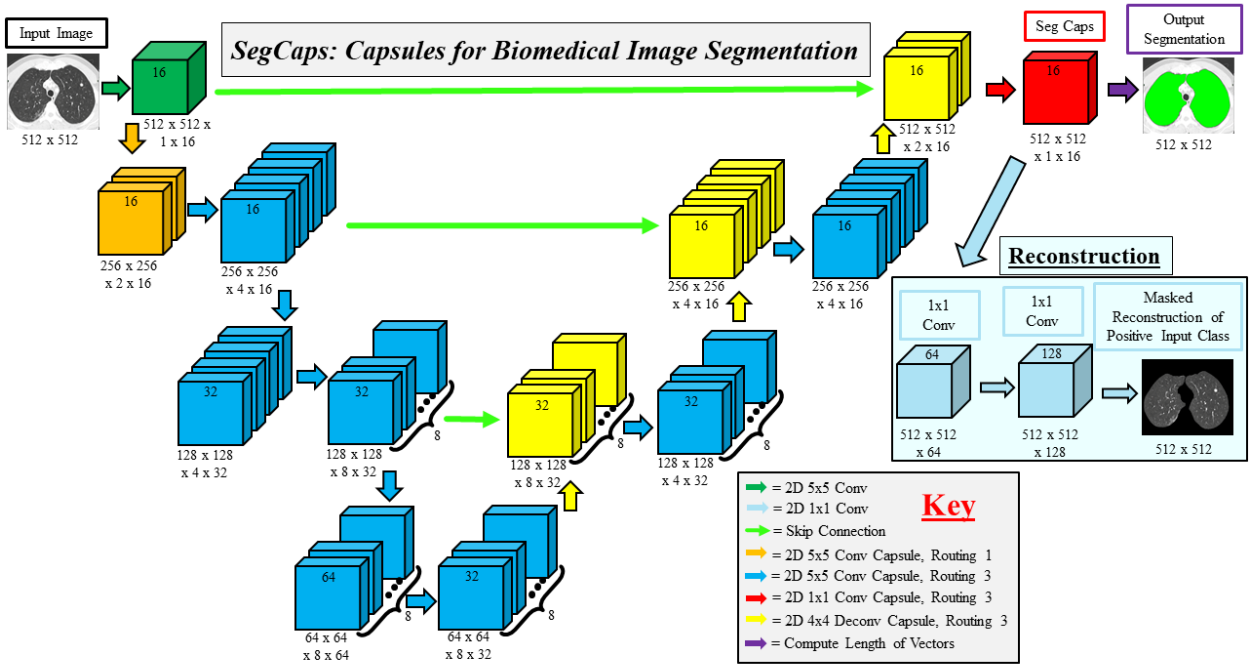


Figure 3.2: The proposed ***SegCaps*** architecture for object segmentation.

3.2 *SegCaps*: Capsules for Object Segmentation

In the following section, we describe the formulation of our *SegCaps* architecture. As illustrated in Figure 2, the input to our *SegCaps* network is a large image (e.g. 512×512 pixels), in this case, a slice of a CT Scan. The image is passed through a 2D convolutional layer which produces 16 feature maps of the same spatial dimensions. This output forms our first set of capsules, where we have a single capsule type with a grid of 512×512 capsules, each of which is a 16 dimensional vector. This is then followed by our first convolutional capsule layer. In the following, we generalize the process of our convolutional capsules and routing to any given layer ℓ in the network.

At layer ℓ, there exists a set of capsule types

$$T^{\ell} = \{t_1^{\ell}, t_2^{\ell}, ..., t_n^{\ell} \mid n \in \mathbb{N}\}. \tag{3.1}$$

For every $t_i^{\ell} \in T^{\ell}$, there exists an $h^{\ell} \times w^{\ell}$ grid of z^{ℓ}-dimensional child capsules,

$$C = \{\boldsymbol{c}_{1,1}, ..., \boldsymbol{c}_{1,w^{\ell}}, ..., \boldsymbol{c}_{h^{\ell},1}, ..., \boldsymbol{c}_{h^{\ell},w^{\ell}}\}, \tag{3.2}$$

where $h^{\ell} \times w^{\ell}$ is the spatial dimensions of the output of layer $\ell - 1$. At the next layer of the network, $\ell + 1$, there exists a set of capsule types

$$T^{\ell+1} = \{t_1^{\ell+1}, t_2^{\ell+1}, ..., t_m^{\ell+1} \mid m \in \mathbb{N}\}. \tag{3.3}$$

And for every $t_j^{\ell+1} \in T^{\ell+1}$, there exists an $h^{\ell+1} \times w^{\ell+1}$ grid of $z^{\ell+1}$-dimensional parent capsules,

$$P = \{\boldsymbol{p}_{1,1}, ..., \boldsymbol{p}_{1,w^{\ell+1}}, ..., \boldsymbol{p}_{h^{\ell+1},1}, ..., \boldsymbol{p}_{h^{\ell+1},w^{\ell+1}}\}, \tag{3.4}$$

where $h^{\ell+1} \times w^{\ell+1}$ is the spatial dimensions of the output of layer ℓ.

In convolutional capsules, for every parent capsule type $t_j^{\ell+1} \in T^{\ell+1}$, every parent capsule $\boldsymbol{p}_{x,y} \in P$ receives a set of "prediction vectors", $\{\hat{\boldsymbol{u}}_{x,y|t_1^{\ell}}, \hat{\boldsymbol{u}}_{x,y|t_2^{\ell}}, ..., \hat{\boldsymbol{u}}_{x,y|t_n^{\ell}}\}$, one for each capsule type in T^{ℓ}. This set of prediction vectors is defined as the matrix multiplication between a learned transformation matrix for the given parent capsule type, $M_{t_j^{\ell+1}}$, and the sub-grid of child capsules outputs, $U_{x,y|t_i^{\ell}}$, within a user-defined kernel centered at position (x,y) in layer ℓ; hence

$$\hat{\boldsymbol{u}}_{x,y|t_i^{\ell}} = M_{t_j^{\ell+1}} \cdot U_{x,y|t_i^{\ell}}, \quad \forall \ t_i^{\ell} \in T^{\ell}. \tag{3.5}$$

Explicitly, each $U_{x,y|t_i^{\ell}}$ has shape $k_h \times k_w \times z^{\ell}$, where $k_h \times k_w$ are the dimensions of the user-defined

kernel, for all capsule types $t_i^\ell \in T^\ell$. Each $M_{t_j^{\ell+1}}$ has shape $k_h \times k_w \times z^\ell \times z^{\ell+1}$. Thus, we can see each $\hat{\boldsymbol{u}}_{x,y|t_i^\ell}$ is an $z^{\ell+1}$-dimensional vector, since these will be used to form our parent capsules. In practice, we solve for all parent capsule types simultaneously by defining M to have shape $k_h \times k_w \times z^\ell \times \mid T^{\ell+1} \mid \times z^{\ell+1}$, where $\mid T^{\ell+1} \mid$ is the number of parent capsule types in layer $\ell+1$. Note, as opposed to CapsNet, we are sharing transformation matrices across members of the grid (i.e. each $M_{t_j^{\ell+1}}$ does not depend on the spatial location (x,y)), as the same transformation matrix is shared across all spatial locations within a given capsule type, similar to how convolutional kernels scan an input feature map. This is one way our method can exploit parameter sharing to dramatically cut down on the total number of parameters to be learned. The values of these transformation matrices for each capsule type in a layer are learned via the backpropagation algorithm with a supervised loss function.

Algorithm 1 Locally-Constrained Dynamic Routing.

procedure ROUTING($\hat{\boldsymbol{u}}_{x,y|t_i^\ell}$, d, ℓ, x, y)
  for all capsule types t_i^ℓ at position (x,y) and capsule type $t_j^{\ell+1}$ at position (x,y): $b_{t_i^\ell|x,y} \leftarrow 0$.
  for d iterations **do**
    for all capsule types t_i^ℓ at position (x,y): $\mathbf{r}_{t_i^\ell} \leftarrow \texttt{softmax}(\mathbf{b}_{t_i^\ell})$ ▷ `softmax` computes Eq. 3.7
    for all capsule types $t_j^{\ell+1}$ at position (x,y): $\boldsymbol{p}_{x,y} \leftarrow \sum_n r_{t_i^\ell|x,y} \hat{\boldsymbol{u}}_{x,y|t_i^\ell}$
    for all capsule types $t_j^{\ell+1}$ at position (x,y): $\boldsymbol{v}_{x,y} \leftarrow \texttt{squash}(\boldsymbol{p}_{x,y})$ ▷ `squash` computes Eq. 3.8
    for all capsule types t_i^ℓ and all capsule types $t_j^{\ell+1}$: $b_{t_i^\ell|x,y} \leftarrow b_{t_i^\ell|x,y} + \hat{\boldsymbol{u}}_{x,y|t_i^\ell} . \mathbf{v}_{x,y}$
  return $\mathbf{v}_{x,y}$

To determine the final input to each parent capsule $\boldsymbol{p}_{x,y} \in P$, where again P is the grid of parent capsules for parent capsule type $t_j^{\ell+1} \in T^{\ell+1}$, we compute the weighted sum over these "prediction vectors" as,

$$\boldsymbol{p}_{x,y} = \sum_n r_{t_i^\ell|x,y} \hat{\boldsymbol{u}}_{x,y|t_i^\ell}, \tag{3.6}$$

where $r_{t_i^\ell|x,y}$ are the routing coefficients determined by the dynamic routing algorithm, and each member of the grid (x,y) has a unique routing coefficient. These routing coefficients are computed

by a "routing softmax",

$$r_{t_i^\ell|x,y} = \frac{\exp(b_{t_i^\ell|x,y})}{\sum_{t_j^{\ell+1}} \exp(b_{t_i^\ell|t_j^{\ell+1}})}, \tag{3.7}$$

whose initial logits, $b_{t_i^\ell|x,y}$ are the log prior probabilities that prediction vector $\hat{\boldsymbol{u}}_{x,y|t_i^\ell}$ should be routed to parent capsule $\boldsymbol{p}_{x,y}$. Note that the $\sum_{t_j^{\ell+1}}$ term is across parent capsule types in $T^{\ell+1}$ for each (x,y) location.

Our method differs from the dynamic routing implemented by Sabour et al. [105] in two ways. First, we locally constrain the creation of the prediction vectors. Second, we only route the child capsules within the user-defined kernel to the parent, rather than routing every single child capsule to every single parent. The output capsule is then computed using a non-linear squashing function

$$\mathbf{v}_{x,y} = \frac{||\boldsymbol{p}_{x,y}||^2}{1+||\boldsymbol{p}_{x,y}||^2} \frac{\boldsymbol{p}_{x,y}}{||\boldsymbol{p}_{x,y}||}, \tag{3.8}$$

where $\mathbf{v}_{x,y}$ is the vector output of the capsule at spatial location (x,y) and $\boldsymbol{p}_{x,y}$ is its final input. Lastly, the agreement is measured as the scalar product,

$$a_{x,y|t_i^\ell} = \mathbf{v}_{x,y} \cdot \hat{\boldsymbol{u}}_{x,y|t_i^\ell}. \tag{3.9}$$

The pseudocode for this locally-constrained dynamic routing is summarized in Algorithm 1. A final segmentation mask is created by computing the length of the capsule vectors in the final layer and assigning the positive class to those whose magnitude is above a threshold, and the negative class otherwise.

3.2.1 Introducing Deconvolutional Capsules

In order to form a deep encoder-decoder network, we introduce the concept of "deconvolutional" capsules. These are similar to the locally-constrained convolutional capsules; however, the prediction vectors are now formed using the transpose of the operation previously described. Note that the dynamic routing of these differently-formed prediction vectors still occurs in the exact same way, so we will not re-describe that part of the operation.

The set of prediction vectors for deconvolutional capsules are defined again as the matrix multiplication between a learned transformation matrix, $M_{t_j^{\ell+1}}$, for a given parent capsule type $t_j^{\ell+1} \in T^{\ell+1}$, and the sub-grid of child capsules outputs, $W_{x,y|t_i^\ell}$ for each capsule type in $t_i^\ell \in T^\ell$, within a user-defined kernel centered at position (x,y) in layer ℓ. However, in deconvolutional capsules, we first need to reshape our child capsule outputs following the fractional striding formulation used by Long et al. [79]. This allows us to effectively upsample the height and width of our capsule grids by the scaling factor chosen. For each member of the grid, we can then form our prediction vectors again by

$$\hat{\boldsymbol{w}}_{x,y|t_i^\ell} = M_{t_j^{\ell+1}} \cdot W_{x,y|t_i^\ell}, \quad \forall \ t_i^\ell \in T^\ell. \tag{3.10}$$

Thus, we have each $\hat{\boldsymbol{w}}_{x,y|t_i^\ell}$ as a $z^{\ell+1}$-dimensional vector, and is input to the dynamic routing algorithm to form our parent capsules. As before, in practice we solve for all parent capsule types simultaneously by defining M to have shape $k_h \times k_w \times z^\ell \times \mid T^{\ell+1} \mid \times z^{\ell+1}$, where $\mid T^{\ell+1} \mid$ is the number of parent capsule types in layer $\ell+1$. Here, we still sharing transformation matrices across members of the grid (i.e. each $M_{t_j^{\ell+1}}$ does not depend on the spatial location (x,y)), similar to how transposed convolutional kernels scan an input feature map.

3.2.2 *Extending Reconstruction Regularization to Segmentation*

As a method of regularization, we extend the idea of reconstructing the input to promote a better embedding of our input space. This forces the network to not only retain all necessary information about a given input, but also encourages the network to better represent the full distribution of the input space, rather than focusing only on its most prominent modes relevant to the desired task. Since we only wish to model the distribution of the positive input class and treat all other pixels as background, we mask out segmentation capsules which do not belong to the positive class and reconstruct a similarly masked version of the input image. We perform this reconstruction via a three layer 1×1 convolutional network, then compute a mean-squared error (MSE) loss between only the positive input pixels and this reconstruction. More explicitly, we formulate this problem as

$$R^{x,y} = I^{x,y} \times S^{x,y} \mid S^{x,y} \in \{0,1\}, \text{ and} \tag{3.11}$$

$$\mathscr{L}_R = \frac{\gamma}{X \times Y} \sum_{x}^{X} \sum_{y}^{Y} \|R^{x,y} - O_r^{x,y}\|, \tag{3.12}$$

where $\mathscr{L}_R$ is the supervised loss for the reconstruction regularization, γ is a weighting coefficient for the reconstruction loss, $R^{x,y}$ is the reconstruction target pixel, $I^{x,y}$ is the image pixel, $S^{x,y}$ is the ground-truth segmentation mask value, and $O_r^{x,y}$ is the output of the reconstruction network, each at pixel location (x,y), respectively, and X and Y are the width and height, respectively, of the input image. An ablation study of the contribution of this regularization is included in Section 3.4. The total loss is the summation of this reconstruction loss and a weighted binary cross-entropy (BCE) loss for the segmentation output, weighted by the foreground/background pixel balance of each training set respectively.

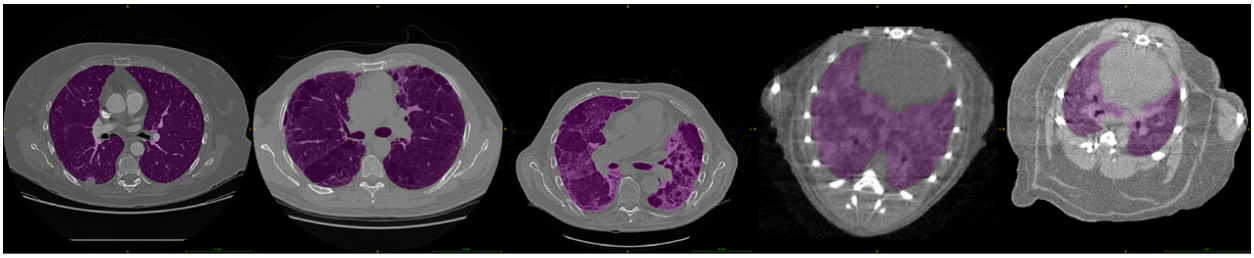


Figure 3.3: Example scans with ground-truth masks (magenta) for each of the five datasets in this study.

3.3 Pathological Lung Segmentation Experiments & Results

3.3.1 Pathological Lung Datasets

Experiments were conducted on five pathological lung datasets, obtained from both clinical and pre-clinical subjects, containing nearly 2000 CT scans, with annotations by expert radiologists. An example typical scan with ground-truth from each dataset is shown in Figure 3.3. The three clinical and two pre-clinical (mice) datasets analyzed are as follows:

- *The Lung Image Database Consortium and Image Database Resource Initiative* [6], abbreviated as **LIDC-IDRI**, contains 885 annotated CT scans of lung cancer screening patients collected from seven academic centers and eight medical imaging companies.
- *The Lung Tissue Research Consortium database* [61], abbreviated as **LTRC**, contains 545 annotated CT scans, with most donor subjects having interstitial fibrotic lung disease or chronic obstructive pulmonary disease (COPD).
- *The Multimedia Database of Interstitial Lung Diseases* [22], abbreviated as **UHG**, built at the University Hospitals of Geneva contains 214 annotated CT scans of patients affected with

one of the 13 histological diagnoses of interstitial lung disease (ILD).

- *The TB-Smoking dataset* collected at Johns Hopkins University, abbreviated as **JHU-TBS**, contains 108 annotated CT scans of mice subjects affected with tuberculosis (TB) and exposed to smoke inhalation.

- *The TB dataset* also collected at Johns Hopkins University, abbreviated as **JHU-TB**, contains 208 annotated CT scans of mice subjects affected with TB undergoing experimental treatment.

In total, 1960 CT scans were annotated in this study. Each dataset was treated completely separate, as each offers unique challenges to automated segmentation algorithms. Ten-fold cross-validation was performed for training all algorithms, with 10% of training data left aside for validation and early-stopping. The mean and standard deviation (std) across the 10-folds for each dataset is presented for two key metrics, namely the 3D Dice similarity coefficient (Dice) and 3D Hausdorff distance (HD) computer for each 3D CT scan.

3.3.2 Implementation Details of SegCaps

All algorithms, namely *U-Net*, *Tiramisu*, *P-HNN*, our three-layer baseline capsule segmentation network, and ***SegCaps*** are all implemented using Keras [19] with TensorFlow [1]. The *U-Net* architecture is implemented exactly as described in the original paper by Ronneberger et al. [102]. *P-HNN* was implemented based on their official Caffe code, including individual layer-specific learning rate multipliers and kernel initialization. However, we removed the layer-specific learning rate and changed the kernel initialization to Xavier to match the other networks and achieve much better results. *Tiramisu* follows the highest performing model presented in [58], namely *FC-DenseNet103*. To remain consistent, since pre-trained models are not available for our custom-designed *SegCaps*, and to better see the performance of each individual method under different

amounts of training data and pathologies present, no pre-trained weights were used to initialize any of the models; instead, all were trained from scratch on each dataset investigated. It can be reasonably assumed based on previous studies that pre-training on large datasets such as ImageNet would improve the performance of all models. A weighted-BCE loss is used for the segmentation output of all networks, with weights determined by the foreground/background pixel balance of each training set respectively. For the capsule network, the reconstruction output loss is computed via the masked-MSE described in Section 3.2.2. All possible experimental factors are controlled between different networks; all networks are trained from scratch, using the same data augmentation methods (scale, flip, shift, rotate, elastic deformations, and random noise) and Adam optimization [64] with an initial learning rate of 0.00001. A batch size of 1 is chosen for all experiments to match the original *U-Net* implementation. The learning rate is decayed by a factor of 0.05 upon validation loss stagnation for $50,000$ iterations and early-stopping is performed with a patience of $250,000$ iterations based on validation 2D Dice scores. Positive/negative pixels were set in the segmentation masks based on a set threshold of 0.5 on the networks' output score maps. All code is made publicly available. [1]

3.3.3 Pathological Lung Segmentation Results

The final quantitative results of these experiments to perform lung segmentation from pathological CT scans are shown in Tables 3.1 - 3.5. Table 3.1 shows results on the LIDC-IDRI dataset, the largest of the three clinical datasets with typically the least severe pathology present on average compared to the other two clinical datasets. Table 3.2 shows results on the LTRC dataset, a large dataset with large amounts of ILD and COPD pathology present. Table 3.3 shows results on the UHG dataset, perhaps the most challenging of the three clinical datasets, both due to its relatively smaller size and

[1] https://github.com/lalonderodney/SegCaps

the severe average amount of pathology present in patients scanned. Table 3.4 shows results on the JHU-TBS dataset, and provides the first fully-automated deep learning based segmentation results presented in the literature for lung segmentation on pre-clinical subjects. Table 3.5 shows results on the JHU-TB dataset, a larger but more challenging dataset of mouse subjects with typically more severe pathology present than the JHU-TBS dataset.

Table 3.1: Experimental results on 885 CT scans from the LIDC-IDRI database [6], measured by 3D Dice Similarity Coefficient and Hausdorff Distance (HD).

Method	Dice (%± std)	HD (*mm*± std)
U-Net [102]	96.06 ± 2.40	41.211 ± 9.109
Tiramisu [58]	94.40 ± 3.66	42.205 ± 15.210
P-HNN [46]	95.64 ± 2.92	41.775 ± 13.866
SegCaps	$\mathbf{96.98 \pm 0.36}$	$\mathbf{30.764 \pm 2.793}$

Table 3.2: Experimental results on 545 CT scans from the LTRC database [61], measured by 3D Dice Similarity Coefficient and Hausdorff Distance (HD).

Method	Dice (%± std)	HD (*mm*± std)
U-Net [102]	95.52 ± 2.80	37.625 ± 6.831
Tiramisu [58]	95.41 ± 2.08	43.969 ± 14.869
P-HNN [46]	95.46 ± 3.93	33.835 ± 9.596
SegCaps	$\mathbf{96.91 \pm 2.24}$	$\mathbf{26.295 \pm 3.806}$

Table 3.3: Experimental results on 214 CT scans from the UHG database [22], measured by 3D Dice Similarity Coefficient and Hausdorff Distance (HD).

Method	Dice (%± std)	HD (*mm*± std)
U-Net [102]	88.10 ± 1.84	44.303 ± 34.148
Tiramisu [58]	87.67 ± 1.38	61.227 ± 54.096
P-HNN [46]	88.64 ± 0.64	43.698 ± 24.026
SegCaps	$\mathbf{88.92 \pm 0.66}$	$\mathbf{37.171 \pm 23.223}$

Table 3.4: Experimental results on 108 CT scans from the JHU-TBS database, measured by 3D Dice Similarity Coefficient and Hausdorff Distance (HD).

Method	Dice (%± std)	HD (*mm*± std)
U-Net [102]	90.38 ± 3.86	7.593 ± 0.886
Tiramisu [58]	86.45 ± 5.76	7.428 ± 1.337
P-HNN [46]	88.81 ± 6.81	7.517 ± 1.896
SegCaps	$\mathbf{93.35 \pm 0.95}$	$\mathbf{4.367 \pm 1.367}$

Table 3.5: Experimental results on 208 CT scans from the JHU-TB database, measured by 3D Dice Similarity Coefficient and Hausdorff Distance (HD).

Method	Dice (%± std)	HD (*mm*± std)
U-Net [102]	76.26 ± 9.51	$\mathbf{24.295 \pm 14.684}$
Tiramisu [58]	79.99 ± 6.24	24.647 ± 11.629
P-HNN [46]	80.11 ± 7.46	26.597 ± 16.168
SegCaps	$\mathbf{80.91 \pm 5.27}$	26.021 ± 10.260

The results of these experiments show ***SegCaps*** consistently outperforms all other compared state-of-the-art approaches in terms of the commonly measured metrics, Dice and HD. Additionally, *SegCaps* achieves this while only using a fraction of the total parameters of these much larger networks. The proposed ***SegCaps*** architecture contains 95.4% fewer parameters than *U-Net*, 90.5% fewer than *P-HNN*, and 85.1% fewer than *Tiramisu*. A comparison with similarly sized version of these other networks is shown in Section 3.4.2. As a brief note in regardless to the discrepancy in results for *P-HNN* between our study and those in the original work, this can be explained by several factors: the original work i) used ImageNet pre-trained models, ii) selected a carefully chosen subset (73 scans) of the UHG dataset, iii) trained and tested models using all datasets combined in the cross-validation splits, and iv) changed the segmentation threshold based on validation scores.

Qualitative results for typical samples from all datasets are shown in Figure 3.4. As can be

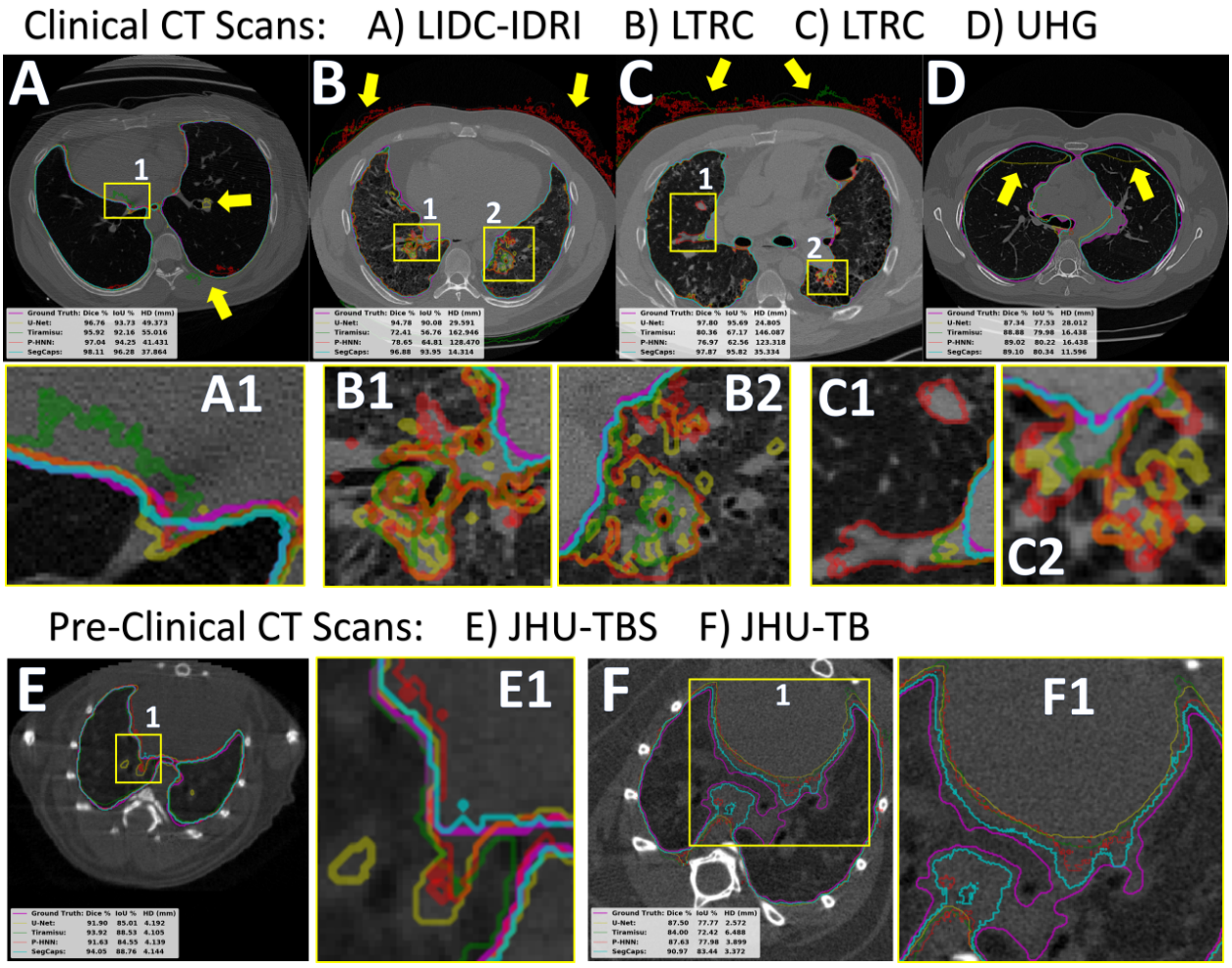


Figure 3.4: Qualitative results on the five datasets analyzed, with quantitative results presented in the lower-left corner of each sub-figure. The top row (A-D) are results on clinical (human) scans; the bottom row (E-F) are results on pre-clinical (mice) scans. It can be noticed that the CNN-based methods' typical failure cases, shown by the yellow arrows and boxed-in regions, are where the pixel intensities (Hounsfield units) are far from the class mean (i.e. high values within the lung regions or low values outside the lung regions). The yellow-boxed regions, with corresponding letters and numbers, are enhanced to more easily see the result contours. Best viewed online in color.

seen in these qualitative examples, *SegCaps* achieves higher results by not falling into the typical segmentation failure causes, namely over-segmentation and segmentation-leakage. These qualitative examples are supported by our quantitative findings where over-segmentation is best captured by

the HD metric and segmentation-leakages are best captured by the Dice metric.

Further, we investigate how different capsule vectors in the final segmentation capsule layer are representing different visual attributes. Figure 3.5 shows three selected visual attributes (each row) out of the sixteen (dimension of final capsule segmentation vector) across different perturbation values of the vectors ranging from -0.25 to +0.25 (each column) for an example clinical and pre-clinical scan. We observe that regions with different textural properties (i.e., small and large homogeneous) are progressively captured by the different dimensions of the capsule segmentation vectors.

3.4 Ablation Studies on Key Components of *SegCaps*

In the following subsections, we investigate the role of the deeper encoder-decoder network structure enabled by the introduction of our deconvolutional capsules, the effect of the reconstruction regularization, the optimal number of dynamic routing iterations to perform, and the relative efficiency of parameter use with similarly-sized versions of all studied networks. The UHG dataset is perhaps the most challenging of the three clinical datasets in our study, both due to its relatively smaller size and the average amount of pathology present in patients scanned. As seen in Tables 3.1 - 3.3, results on all metrics are significantly lower for this challenging dataset. For those reasons, and the lower performance scores leading to bigger differences between approaches, as well as the dataset being publicly available, we chose this dataset for running our ablation experiments.

3.4.1 SegCaps Network Structure/Deconvolutional Capsules

The original CapsNet introduced by Sabour et al. [105] was a simple three layer network, consisting of a single convolutional layer, a primary capsule layer (convolutional layer with a reshape function),

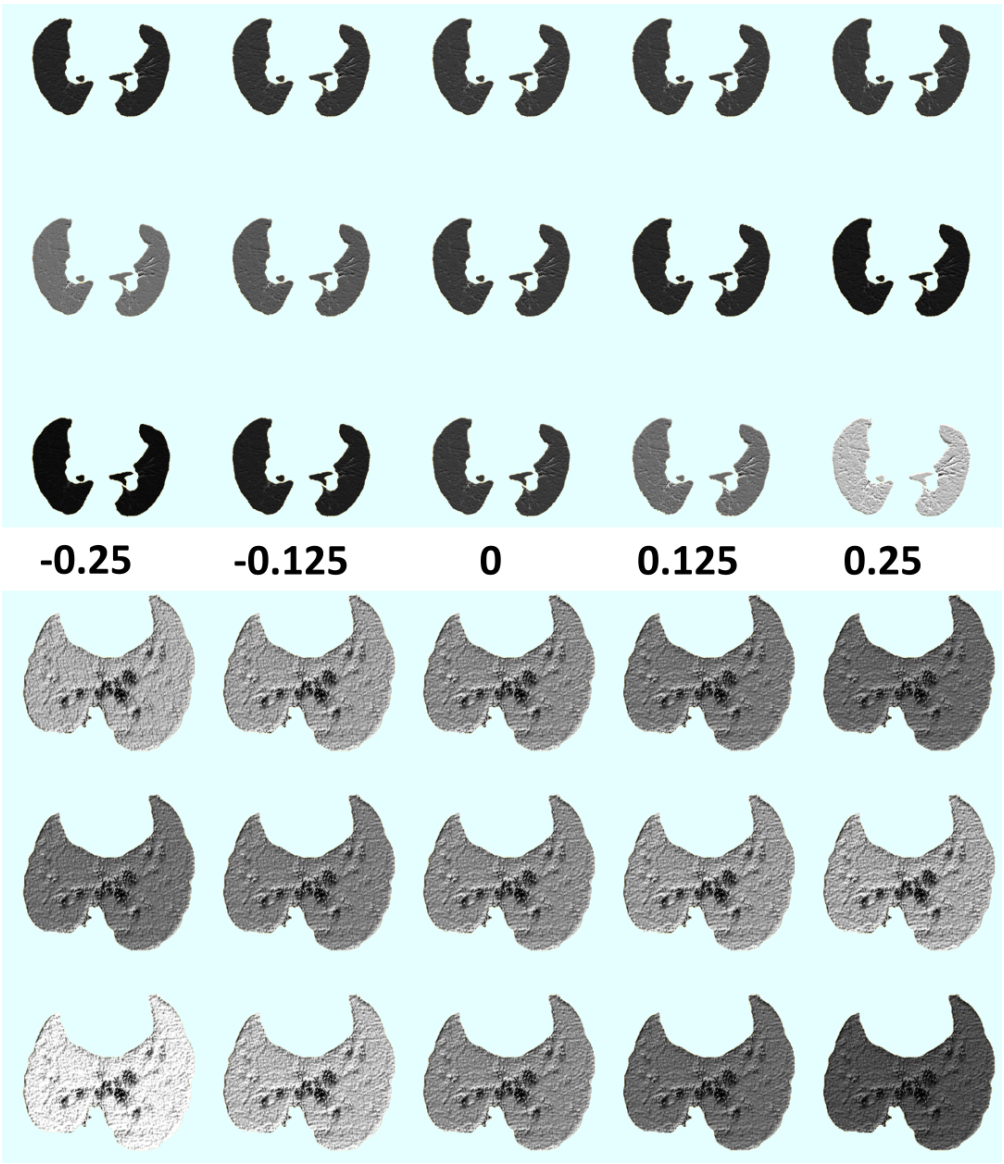


Figure 3.5: Reconstructions of selected capsule vectors (rows) under different perturbations from -0.25 – 0.25 (columns). The top three rows are reconstructions of a scan slice from the clinical LTRC dataset, while the bottom three are from the pre-clinical JHU-TB dataset. These results demonstrate that different dimensions of the capsule vectors are in fact learning different attributes of the lung tissue being segmented.

and a fully-connected capsule layer. This network achieved remarkable results for its size, beating the state-of-the-art on MNIST and performing well on CIFAR10. In our initial efforts for this study,

we attempted to apply this network to the task of segmentation, however, the fully-connected capsule layer was far too memory intensive to make this approach viable with our 512×512 2D slices of CT scans. After introducing the locally-constrained dynamic routing and transformation matrix sharing, we then created a network nearly identical to the original CapsNet with the fully-connected capsule layer swapped out for our locally-constrained version. A diagram of this network is shown in Figure 3.1. The results of this network on the UHG dataset is shown in Table 3.6. As one might expect, swapping out a layer which is fully-connected in space for one which is locally-connected dramatically hurt the performance for a task which relies on global information (*i.e.* determining lung tissue/air from non-lung tissue, bone, etc.). This motivated the introduction of the "deconvolutional" capsule layer which allows for the creation of deep encoder-decoder networks, and thus the recovery of global information, retention of local information, and the parameter savings of locally-constrained capsules.

Table 3.6: Comparing the deeper encoder-decoder network structure *SegCaps* enabled by our proposed deconvolutional capsules, versus a network designed to be as similar as possible to CapsNet [105] (*Baseline SegCaps*), abbreviated in table as Base-Caps.

Method	Dice (%$\pm$ std)	HD (*mm*$\pm$ std)
Base-Caps	75.97 ± 4.60	352.582 ± 133.451
SegCaps	$\mathbf{88.92 \pm 0.66}$	$\mathbf{37.171 \pm 23.223}$

3.4.2 Comparing Similar Parameter Usage

Shown in Tables 3.7– 3.8, we investigate the number of parameters in the proposed *SegCaps*, *U-Net*, *Tiramisu* and *P-HNN*, as well as down-scaled versions of *U-Net*, *Tiramisu*, and *P-HNN*. *U-Net* and *P-HNN* are scaled down by dividing the number of feature maps per layer by a constant factor, $k = 4.68$ and $k = 3.2$ respectively, and *Tiramisu* is scaled down by using the lighter FC-DenseNet56 purposed in the original work by Jégou et al. [58]. When the parameters of *U-Net* and *P-HNN*

Table 3.7: Number of parameters for each of the networks examined in this study. The percentage of less parameters (Percent Less) is measured relative to the number of parameters in *U-Net*.

Method	Parameters	Percent Less
U-Net	31.0 M	0.00 %
P-HNN	14.7 M	52.58 %
Tiramisu	9.4 M	69.68 %
Baseline SegCaps	1.7 M	94.52 %
SegCaps	**1.4 M**	**95.48 %**

are scaled down to roughly the same number of parameters as *SegCaps*, these models perform comparatively worse, as shown in Table 3.8, providing evidence that *SegCaps* is able to make better use of the parameters available to it than its CNN counterparts. *Tiramisu-56* is a minor exception to this trend as its Dice score remained similar while the HD only fell slightly from *Tiramisu-103*. The reason for this is most likely because *Tiramisu-56* was carefully engineered to achieve the highest possible accuracy with few parameters while the addition of dense connections has been shown to make far better use of parameters than standard non-dense CNNs [53]. However, as can be see in Table 3.8, when all networks have roughly the same number of parameters, *SegCaps* outperforms all other methods.

3.4.3 Reconstruction Regularization Performance

The idea of reconstructing the input as a method of regularization was used in CapsNet by Sabour et al. [105]. The theory behind this technique and the regularization effect it introduces is similar in nature to the problem of "mode collapse" in generative adversarial networks (GANs). When training a generative neural network for a specific task through the backpropagation algorithm, the model "collapses" to focusing on only the most prevalent modes in the data distribution. A similar phenomenon occurs when you train a discriminative network for a specific task, the model

Table 3.8: Experimental results on the UHG dataset using downscaled version of *U-Net* and *Tiramisu* to roughly equal the same number of parameters (1.4 M) as *SegCaps*. The value of k (number of feature maps per layer reduction factor) for *U-Net* and *P-HNN* is included in parentheses.

Method	Dice (%± std)	HD (*mm*± std)
U-Net (orig.)	88.10 ± 1.84	44.303 ± 34.148
U-Net (4.68)	87.57 ± 2.80	62.006 ± 62.693
Tiramisu-103	87.67 ± 1.38	61.227 ± 54.096
Tiramisu-56	87.68 ± 0.96	67.913 ± 36.190
P-HNN (orig.)	88.64 ± 0.64	43.698 ± 24.026
P-HNN (3.2)	86.69 ± 1.39	82.223 ± 48.989
SegCaps	$\mathbf{88.92 \pm 0.66}$	$\mathbf{37.171 \pm 23.223}$

"collapses" to only focus on the most discriminative features in the input data and ignores all others. By mapping the capsule vectors back to the input data, this forces the network to pay attention to more relevant features about the input, which might not be as discriminative for the given task, yet still provide some useful information, as evident by the improved results shown in Table 3.9. A similar results can be seen in VEEGAN by Srivastava et al. [114], where they help solve the issue of mode collapse in GANs through a reconstructor network which reverses the action of the generator by mapping from data to noise.

Table 3.9: Examining the effect of the proposed extension of the reconstruction regularization to the task of segmentation.

Method	Dice (%± std)	HD (*mm*± std)
No Recon	88.58 ± 1.03	42.345 ± 21.180
With Recon	$\mathbf{88.92 \pm 0.66}$	$\mathbf{37.171 \pm 23.223}$

3.4.4 Examining Dynamic Routing Iterations

Since the dynamic routing algorithm chosen for this study is an iterative process, we can investigate the optimal number of times to run the routing algorithm per forward pass of the network. In the original work by Sabour et al. [105], they found three iterations to provide the optimal results. As seen in Table 3.10, the number of routing iterations does have an effect on the network's performance, and we find the same result in this study of three iterations being optimal over a set of different numbers of iterations studied.

Table 3.10: Examining the effect of different number of routing iterations (abbreviated as # Iters) per forward pass of *SegCaps*. In $1,3$, one routing iteration is performed when the spatial resolution remains the same and three iterations are performed when the resolution changes.

# Iters	Dice (%$\pm$ std)	HD (*mm*$\pm$ std)
1	88.17 ± 1.23	67.668 ± 58.556
2	88.58 ± 1.03	42.345 ± 21.180
3	$\mathbf{88.92 \pm 0.66}$	$\mathbf{37.171 \pm 23.223}$
4	87.72 ± 1.36	110.901 ± 71.701
$1,3$	88.11 ± 1.13	72.877 ± 54.649

3.5 Applicability of *SegCaps* to Other Data

To demonstrate the extended scope and potential impact of our study, we have performed two additional sets of experiments in object segmentation:

1. Segmenting retinal vessels, containing extremely thin tree-like structures, from retinal angiography video.

2. Testing the affine equivariant properties of *SegCaps* on natural images from PASCAL VOC [34].

The results of these two experiments are shown in Figure 3.6 and Figure 3.7, respectively.

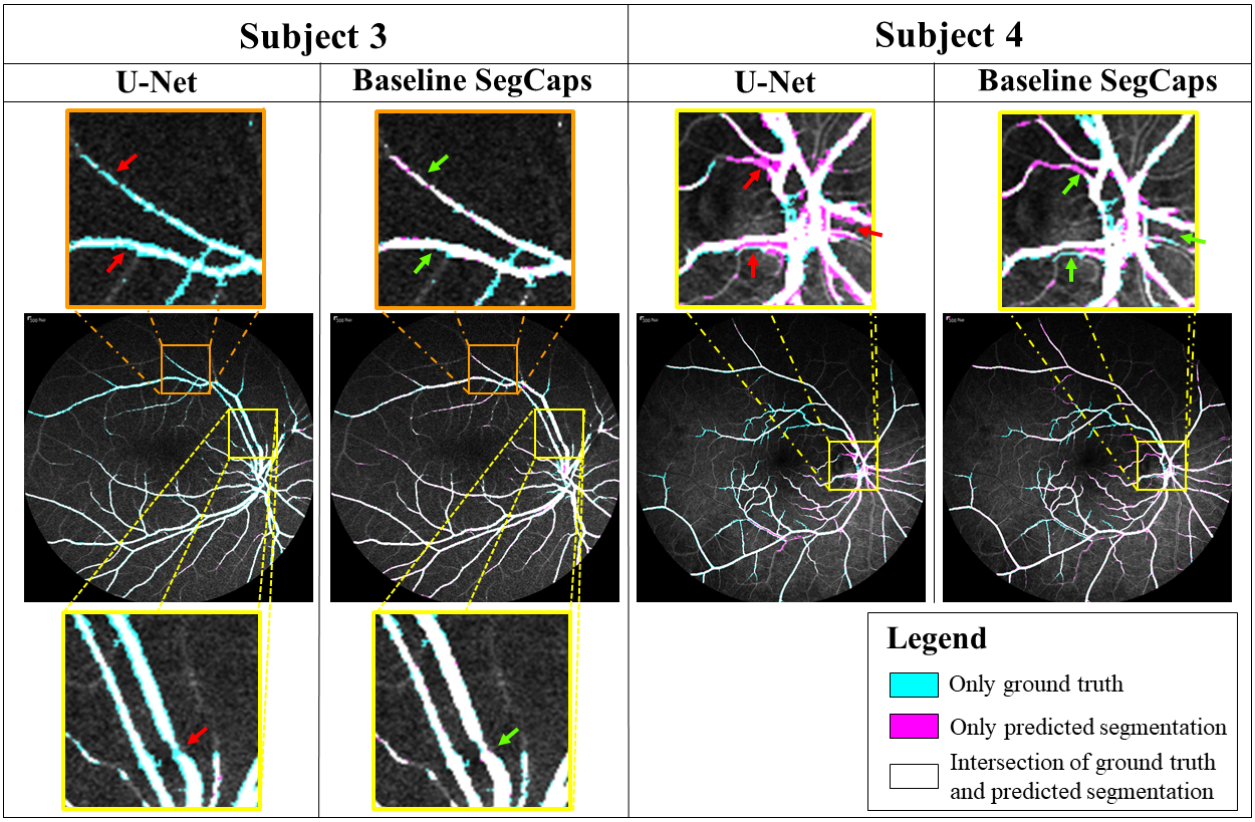


Figure 3.6: Comparing the performance of *U-Net* against the proposed *Baseline SegCaps* network for segmenting thin retinal vessels. The arboriform structure of these vessels can be extremely difficult to segment, especially the thin vessels off the main branches. Where U-Net suffers from both under-segmentation (Subject 3) and over-segmentation (Subject 4) issues, *Baseline Segcaps* performs consistently better.

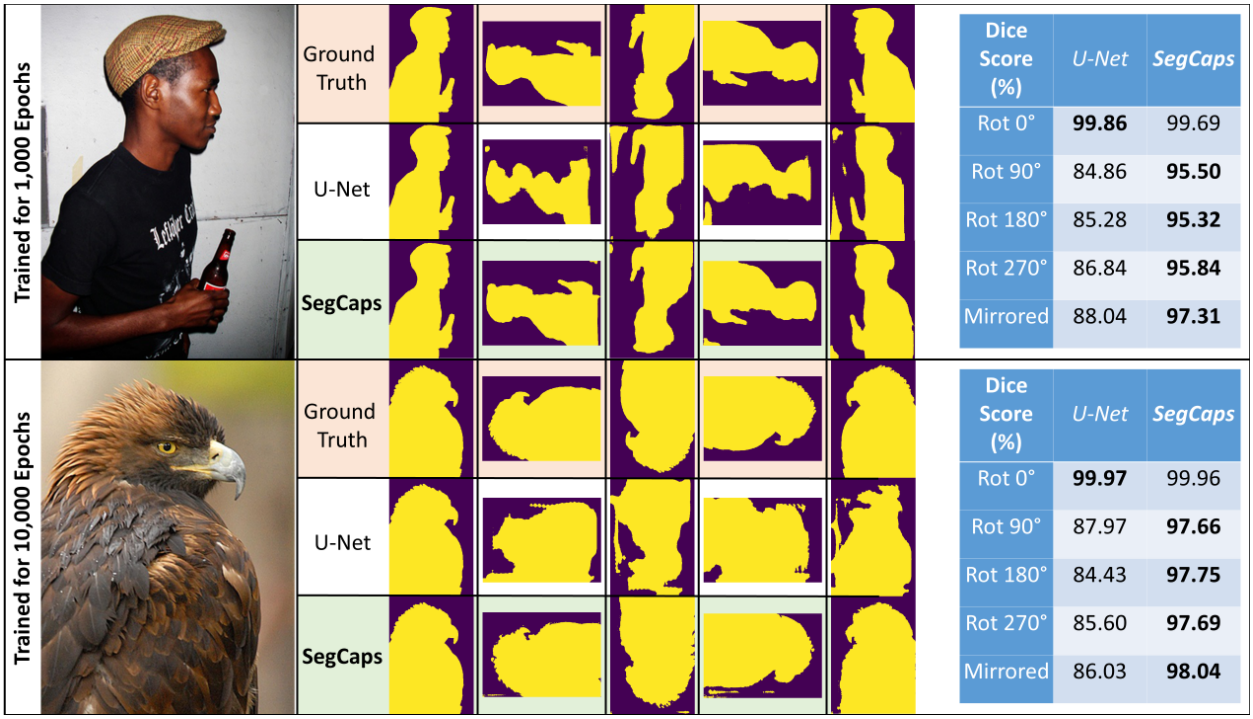
3.5.1 Retinal vessel segmentation

To demonstrate the capabilities of *SegCaps* on other types of object structures, we ran a small-scale experiment on segmenting retinal vessels from fluorescein angiogram videos. The results of this experiment are highlighted in Figure 3.6. Videos of blood flow through retinal vessels was obtained

from 10 subjects, where 10-fold cross-validation was performed to train *U-Net* and our *Baseline SegCaps* model. While *Baseline SegCaps* provides consistently good performance across all subjects, *U-Net* struggles with issues of over-segmentation and under-segmentation, particularly when dealing with thin and crowded vessels. This experiment shows that our capsule-based segmentation method can handle segmenting all manner of objects, from the large lung fields in CT scans to thin tree-like structures in retinal angiography video.

3.5.2 Generalizing to unseen poses in segmentation

In the second experiment, we tested the affine equivariant property of capsule networks on natural images. It has been stated that, due to the affine projections of capsule vectors from children to parents, capsules should be robust to affine transformations on the input, and should in fact be able to generalize to *unseen* poses of target classes. However, no study has formally demonstrated this property. In this experiment, we randomly selected images from the PASCAL VOC dataset which contained only a single foreground object. Both *U-Net* and *SegCaps* were then trained on a single selected image until training accuracy converge to 100%, which occurred around 1000 epochs for both networks. Each network was then tested on 90 degree rotations and the mirroring of the training image. *SegCaps* performed well on nearly all images tested, while *U-Net* performed quite poorly, as can be seen in Figure 3.7. Since *U-Net* has significantly more parameters than *SegCaps*, we also ran experiments at 10000 epochs, long after both networks had converged to 100% training accuracy. This improved the results of *U-Net* on many images; however, there were still significant failure cases, where *SegCaps* did not suffer the same issue. Not only does this show that *SegCaps* is indeed far more robust to affine transformations on the input, a significant issue for CNNs as shown in both this experiment and works such as by Alcorn et al. [3], but also that *SegCaps* converges significantly faster during training that *U-Net*.



Dice Score (%)	*U-Net*	***SegCaps***
Rot 0°	**99.86**	99.69
Rot 90°	84.86	**95.50**
Rot 180°	85.28	**95.32**
Rot 270°	86.84	**95.84**
Mirrored	88.04	**97.31**

Dice Score (%)	*U-Net*	***SegCaps***
Rot 0°	**99.97**	99.96
Rot 90°	87.97	**97.66**
Rot 180°	84.43	**97.75**
Rot 270°	85.60	**97.69**
Mirrored	86.03	**98.04**

Figure 3.7: Testing the affine equivariant properties of capsule networks, specifically *SegCaps*, by overfitting on a single image, trained without augmentation, then predicting on transformations of that image.

3.6 Discussions & Conclusion on Capsule-Based Segmentation

We propose a novel deep learning algorithm, called ***SegCaps***, for object segmentation, and showed its efficacy in a challenging problem of pathological lung segmentation from CT scans. The proposed framework is the first use of the recently introduced capsule network architecture and expands it in several significant ways. First, we modify the original dynamic routing algorithm to act locally when routing children capsules to parent capsules and to share transformation matrices across capsules within the same capsule type. These changes dramatically reduce the memory and parameter burden of the original capsule implementation and allows for operating on large image sizes, whereas previous capsule networks were restricted to very small inputs. To compensate

for the loss of global information, we introduce the concept of "deconvolutional capsules" and a deep convolutional-deconvolutional capsule architecture for pixel level predictions of object labels. Finally, we extend the masked reconstruction of the target class as a regularization strategy for the segmentation problem.

Experimentally, ***SegCaps*** produces improved accuracy for lung segmentation on five datasets from clinical and pre-clinical subjects, in terms of Dice coefficient and Hausdorff distance, when compared with state-of-the-art networks *U-Net* [102], *Tiramisu* [58], and *P-HNN* [46]. More importantly, the proposed ***SegCaps*** architecture provides strong evidence that the capsule-based framework can more efficiently utilize network parameters, achieving higher predictive performance while using 95.4% fewer parameters than *U-Net*, 90.5% fewer than *P-HNN*, and 85.1% fewer than *Tiramisu*. To the best of our knowledge, this work represents the largest study in pathological lung segmentation, and the only showing results on pre-clinical subjects utilizing state-of-the-art deep learning methods.

To demonstrate the extended scope and potential impact of our study, we have performed two additional sets of experiments in object segmentation included in Section 3.5 of this study. The results of these experiments, as well as the main body of our study, demonstrate the effectiveness of the proposed capsule-based segmentation framework. This study provides helpful insights into future capsule-based works and provides lung-field segmentation analysis on pre-clinical subjects for the first time in the literature.

CHAPTER 4: A CAPSULE-BASED MEDICAL DIAGNOSIS FRAMEWORK

Related Publications and Patents:

- Rodney LaLonde, Pujan Kandel, Concetto Spampinato, Michael B Wallace, and Ulas Bagci. "Diagnosing Colorectal Polyps in the Wild with Capsule Networks." *ISBI 2020, IEEE International Symposium on Biomedical Imaging*. 2020. [73]
- Pujan Kandel, Rodney LaLonde, Victor Ciofoaia, Michael B Wallace, and Ulas Bagci. "Colorectal Polyp Diagnosis with Contemporary Artificial Intelligence." *Gastrointestinal Endoscopy*, 89(6), AB403. 2019. [60]
- Rodney LaLonde and Ulas Bagci. "Capsules for Image Analysis." *U.S. Patent Application 16/431,387; filed December 5, 2019*.

In this chapter, we expand on the work completed in Chapter 3, by introducing a capsule-average pooling function to create a deep capsule network for diagnosing colorectal polyps. The remainder of this chapter is organized into the following sections: Section 4.1 – A brief overview of colorectal polyp diagnosis; Section 4.2 – The *D-Caps* framework described in detail, including the capsule-average pooling algorithm; Section 4.3 – Experiments conducted, our implementation settings, and their results; Section 4.4 – Ablation studies on the various components of our algorithms to determine the contribution of each aspect of our proposed method to the final results; and Section 4.5 – Discussions and concluding remarks.

4.1 A Brief Overview of Colorectal Polyp Diagnosis

Among all cancer types, colorectal cancer remains one of the leading causes of cancer-related death worldwide, with the lifetime risk of developing colorectal cancer around 1 in 23 in the United States, accounting for roughly 10% of all cases across genders [4]. The gold standard for colorectal cancer diagnosis is based on the biopsy of colon polyps found during screening (colonoscopy). Due to the vast majority of colorectal cancer cases arising from precursor lesions, referred to as polyps, the identification and resection of pre-malignant polyps during colonoscopy has been shown to decrease colorectal cancer incidence by 40 – 60% [12]. However, small and diminutive polyps make up over 90% of polyps detected, with less than half of these classified as pre-malignant, making diagnosis through 'optical biopsy' by colonoscopists difficult.

Colorectal polyps are typically classified into one of three categories: hyperplastic, serrated (comprised of sessile serrated adenomas and traditional serrated adenomas), and adenomas. Example polyps can be seen in Fig. 4.1. Serrated polyps and adenomas are considered premalignant and should be resected during colonoscopy, while hyperplastic polyps are considered benign and can safely be left *in situ*. Unfortunately, existing optical biopsy techniques, cannot currently be recommended in routine clinical practice due to test accuracy and sensitivity falling substantially below recommended levels [26]. Therefore, current standards require taking a sample of the polyp and performing histopathological analysis, a somewhat time-consuming and expensive process. Further, performing polypectomies (i.e., biopsy) on non-premalignant polyps is unnecessary, increases procedure-related risks such as perforation and bleeding, and increases procedure-related costs including the cost of histological analysis for diagnosis. Improvements in colonoscopy and optical biopsy techniques have been developed [27, 33]; however, with increased colonoscopy use causing an increase in detected polyps, expecting endoscopists to perform optical diagnosis during colonoscopy screenings might prove too time-consuming to manage in routine clinical practice.

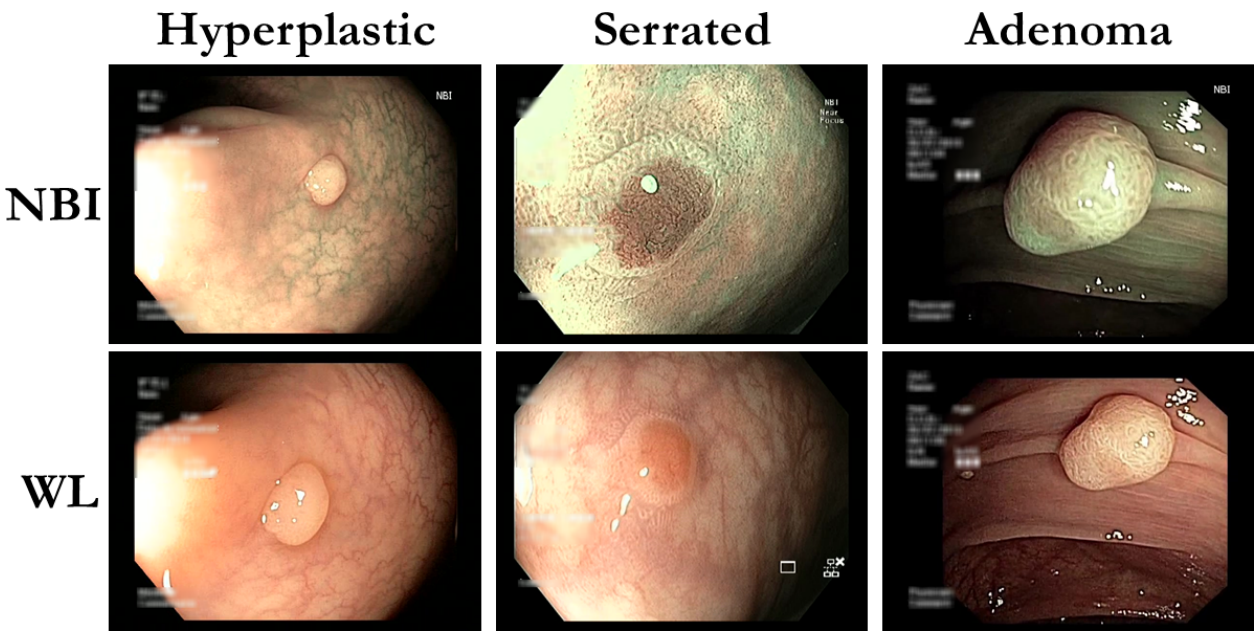


Figure 4.1: Among the most ideal image cases selected from the Mayo Polyp dataset to provide the reader with a visual understanding of the differences between the diagnosis classes and imaging modalities (NBI – narrow-band imaging; WL – white light).

There is a high-expectation for artificial intelligence (AI), particularly deep learning, approaches to be adopted into clinical settings for earlier and more accurate diagnosis of cancers.

Research Gap: Previous academic works have achieved remarkable success in this difficult task, with accuracy scores just exceeding 90% [16, 31]. **However, these methods have been applied to academic datasets which are highly unrealistic compared to a 'real-world' clinical setting.** For example, the most popular dataset in the literature is the ISIT-UMR Multimodal classification dataset [30], containing only 76 polyps. Each polyp is recorded up-close for approximately 30 seconds (nearly 800 videos frames) from multiple angles, modalities, and focus modes. Such time-consuming and ideal videos cannot be expected in more realistic 'in the wild' (i.e., real-world) clinical settings. To address this discrepancy between ideal academic datasets and real-world examples, we performed experiments on the significantly more challenging Mayo Polyp classification dataset, collected at the

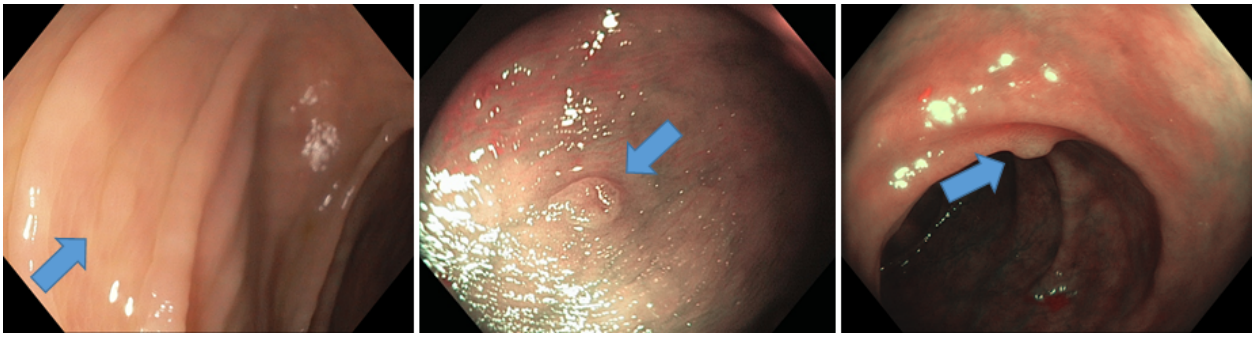


Figure 4.2: Typical cases on real-world ('in-the-wild') polyp diagnosis cases from the Mayo Polyp dataset. Left to right: hyperplastic, serrated, and adenoma, marked by blue arrows.

Mayo Clinic, Jacksonville by [28] with institutional review board approval. A total of 963 polyps from 552 patients were collected, where one image per imaging type of each polyp are chosen by expert interpreters. This dataset is extremely challenging, having only **single images per imaging mode per polyp**, large inter-polyp variation (e.g., scale, skew, illumination), and often only a single imaging mode provided, while also containing far more polyps collected from more patients than all previous AI-driven diagnosis studies in this area. Examples from the Mayo Polyp dataset which are more representative of the typical images are shown in Fig. 4.2, as opposed to the ideal cases handpicked for Fig. 4.1.

To accomplish our task and improve the viability of optical biopsy of colorectal polyps, we design a novel capsule network (D-Caps). As stated in the previous chapters, capsule networks provide equivariance to affine transformations on the input through encoding orientation information in vectorized feature representations. Because of this, *we hypothesize that a capsule network can better model the high intra-class variation present given the relatively limited data in the Mayo Polyp dataset and provide superior results to a deep CNN*. Our method introduces several technical novelties including (i) a novel deep capsule network architecture based on the locally-constrained routing introduced in [71], (ii) a capsule-average pooling (CAP) technique which allows us to

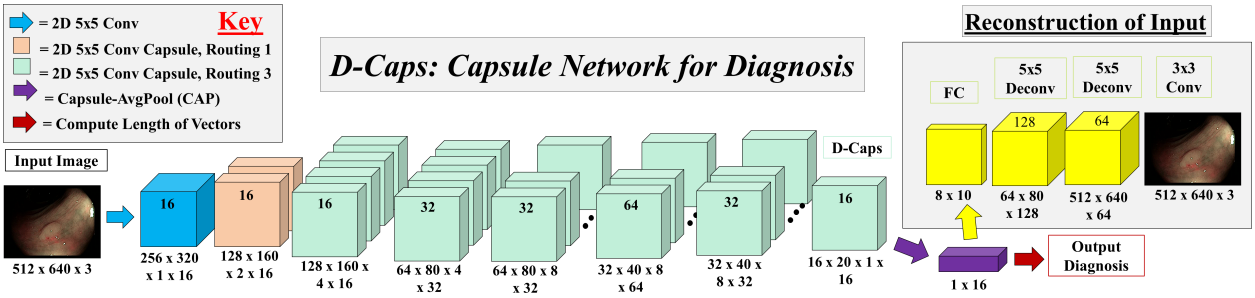


Figure 4.3: D-Caps: Diagnosis capsule network architecture. Routing 1 or 3 refers to the number of routing iterations performed.

perform classification on large image sizes, where the original fully-connected capsules of [105] are far too computationally expensive to fit in GPU memory, and (iii) improves the results over CNNs such as Inceptionv3 (Iv3) [116] employed the previous state-of-the-art [16] by a significant margin, while also reducing the amount of parameters used by as much as 95%. We provide extensive analysis of results stratified across polyp categories, scanner types, imaging modalities, and focus modes to establish a new benchmark on this challenging, unexplored, large-scale dataset and promote future direction into the use of AI-driven colorectal cancer screening systems.

4.2 *D-Caps*: A Diagnosis Capsule Framework

The proposed D-Caps is illustrated in Fig 4.3. Briefly, input to the network is a $512 \times 640 \times 3$ color image taken during colonoscopy screening. This image is sent through an initial convolutional layer which downsamples the image and extracts the basic low-level feature information (edges and corners). This output is reshaped to be treated as a convolutional capsule with a single capsule type, whose feature vectors are then passed to the first convolutional capsule layer, referred to as the primary capsules and a second capsule type is added. All further layers are convolutional capsule layers with locally connected dynamic routing, until the capsule-average pooling layer and

reconstruction sub-network.

In each capsule layer, there are individual capsules which form a grid. Then, at each layer, there are multiple sets of these grids which form the capsule types. Capsules within a lower layer are referred to as child capsules and in a higher layer being routed to as parent capsules. The locally connected dynamic routing works by forming prediction vectors over a kernel of the child capsules centered at the location of the set of parent capsule types. For every parent capsule at a given (x,y) position, a set of prediction vectors are formed via the multiplication between a locally-defined window and a transformation matrix which is shared across the spatial dimension (but not the capsule type dimension). These transformation matrices act analogous to affine transformation in feature space, allowing for a strong notion of equivariance to input features. Once prediction vectors are formed for a given (x,y) location, and therefore set of parent capsules, the modified dynamic routing algorithm then routes all child capsules to all parents capsules only at that given spatial location.

The **capsule-average pooling (CAP) layer** computes the spatial average of capsule activation vectors to reduce the dimensionality of the features. Each capsule type computes an element-wise *mean* across the height and width dimensions of the capsule grid, preserving the length of the capsule vectors in each capsule type. Since, in our application, we are computing a binary classification, we have one capsule type in the final convolutional capsule layer, which transforms to a 1D vector of length k, in our case $k = 16$. More explicitly, if we have n capsule types, each with $h \times w$ grids of capsule vectors of length a, we compute

$$p_a^i = \frac{1}{h \times w} \sum_h \sum_w c_{h,w,a}^i, \forall i \in \{1,2,...,n\}. \tag{4.1}$$

In previous approaches, a fully-connected capsule layer is used to predict the final class-activation vectors. This becomes computationally infeasible with any reasonable sized GPU memory when

working with large-scale images and number of classes. By utilizing our CAP layer, we are able to dramatically increase the size of the images we work with beyond the likes of MNIST, CIFAR-10 and smallNORB. The D-Caps architecture shown in Fig. 4.3 contains only 1.3 million parameters, as compared to 24 million in Inceptionv3, a relative reduction of 95%, while achieving higher performance.

To decide a class score: the magnitude of each vector is computed, where the longest vector is chosen as the prediction. In the case where multiple images of the same polyp were given, the votes for each images are averaged, weighted by the relative confidence of the vote being cast. Reconstruction of the input is then performed via a dense layer followed by two deconvolutions and a final convolution. The reconstruction serves the purpose of providing a learned inverse mapping from output to input, in order to help preserve a better approximation of the distribution of the input space. Without the inverse mapping, the network will be prone to only learn the most common modes in the training dataset. We show in an ablation study this reconstruction significantly helps the accuracy of our approach, which is not possible with a standard CNN that only represents features as scalars.

4.3 Colorectal Polyp Diagnosis Experiments & Results

Experiments were performed on a Mayo Polyp dataset, collected at the Mayo Clinic, Jacksonville by [28] with an institutional review board approval. A total of 552 patients were included in this study with 963 polyps collected. Polyps were collected from both standard and dual-focus colonoscopes. The dual-focus colonoscope contains near and far modes for both white light (WL) and narrow-band imaging (NBI) settings, referred to as WL-N, WL-F, NBI-N, and NBI-F, respectively. Challenging images of each polyp type are chosen by expert interpreters (one per imaging type).

Table 4.1: Classifying **Hyperplastic vs Adenoma polyps** measured by accuracy (acc), sensitivity (sen), and specificity (spe), where -F and -N denote far and near focus, respectively.

Modality	D-Caps			Inceptionv3		
	Acc. %	Sen. %	Spec. %	Acc. %	Sen. %	Spec. %
All Images	63.66	65.26	60.00	54.28	54.83	53.00
All Polyps	65.53	71.12	53.79	56.23	63.18	41.67
NBI	56.69	54.23	61.98	52.49	57.69	41.32
NBI-F	53.37	51.97	57.14	58.65	59.21	57.14
NBI-N	60.95	59.74	64.29	53.33	56.49	44.64
WL	68.81	74.06	57.38	55.41	54.89	56.56
WL-F	72.48	75.63	63.79	55.50	53.75	60.34
WL-N	67.65	70.86	58.49	58.33	58.94	56.60
Near	67.57	70.19	60.66	57.66	63.35	42.62
Far	69.64	73.62	59.02	58.48	63.19	45.90

Three sets of experiments were conducted using stratified 10-fold cross validation. In the first set, images were split into two categories, hyperplastics and adenomas (with serrated adenomas excluded). This is the most common training split in previous studies in this research area. In the second set, the serrated adenomas were included in the adenoma class. This experiment is the most clinically meaningful. Both Adenomas and Serrated polyps needed to be resected during optical biopsy, thus this represents the task equivalent to resect of leave *in-situ*. In the third set, images were split between hyperplastics and serrated adenomas with the adenoma images excluded. The reasoning behind this experiment is Serrated polyps are most frequently mistaken with Hyperplastic polyps and the degree to which a deep learning based approach can separate these classes is of scientific interest.

All networks were trained and tested on a single Titan X GPU using the Keras and TensorFlow frameworks. Both Inceptionv3 and D-Caps were trained from scratch using the Adam optimizer at its default settings. A batch size of 8 was used for Inceptionv3 and 4 for D-Caps due to memory constraints on capsules. The loss function for all networks was a binary cross-entropy. All code for

Table 4.2: Classifying **Hyperplastic vs Adenoma and Serrated polyps** measured by accuracy (acc), sensitivity (sen), and specificity (spe), where -F and -N denote far and near focus, respectively.

Modality	D-Caps			Inceptionv3		
	Acc. %	Sen. %	Spec. %	Acc. %	Sen. %	Spec. %
All Images	59.81	61.39	56.00	51.21	53.49	45.75
All Polyps	60.95	63.19	56.06	48.10	50.35	43.18
NBI	60.36	60.00	61.16	45.27	41.11	54.55
NBI-F	60.09	59.24	62.50	51.17	50.96	51.79
NBI-N	63.59	65.22	58.93	46.54	46.58	46.43
WL	54.39	59.21	43.44	51.88	56.68	40.98
WL-F	55.86	64.02	32.76	59.01	66.46	37.93
WL-N	56.67	58.60	50.94	50.95	58.60	28.30
Near	58.52	60.12	54.10	47.60	51.79	36.07
Far	62.01	67.86	45.90	50.22	55.36	36.07

reproducing experiments are made publicly available. [1]

The results of the three sets of experiments are presented in Tables 4.1 - 4.3. For all experiments, we stratify results at several levels of analysis: All Images presents results for every image present in the dataset, while all other results are a weighted average taken across all votes for a given polyp (and imaging modality) to give a final diagnosis score. Looking at the All Polyps rows, we can see *D-Caps* outperforms *Inceptionv3* in terms of relative accuracy increases of 17%, 27%, and 43% for experiments 1 – 3 (of increasing difficulty) respectively. We also provide some qualitative examples of success and failure cases shown in Fig. 4.4.

4.4 Ablation Studies on Key Components of *D-Caps*

We conducted three rounds of ablation experiments in this study, with results presented at the polyp level. First, we vary the amount of dynamic routing iterations performed inside D-Caps. Second, we

[1] https://github.com/lalonderodney/D-Caps

Table 4.3: Classifying **Hyperplastic vs Serrated polyps** measured by accuracy (acc), sensitivity (sen), and specificity (spe), where -F and -N denote far and near focus, respectively.

Modality	D-Caps			Inceptionv3		
	Acc. %	Sen. %	Spec. %	Acc. %	Sen. %	Spec. %
All Images	60.91	65.00	60.50	51.45	63.64	50.62
All Polyps	58.04	54.55	58.33	40.54	66.67	39.05
NBI	57.85	70.00	56.76	45.63	83.33	43.30
NBI-F	55.00	60.00	54.29	44.90	66.67	43.48
NBI-N	60.00	71.43	57.58	50.00	100.00	46.81
WL	54.14	54.55	54.10	48.08	16.67	50.00
WL-F	52.63	100.00	49.06	41.86	66.67	40.00
WL-N	52.54	50.00	52.83	45.00	33.33	45.95
Near	67.21	57.14	68.52	40.00	100.00	36.17
Far	66.67	60.00	67.21	39.62	66.67	38.00

remove the reconstruction regularization sub-network from the loss formulation. Third, we evaluate D-Caps predictive performance on an 'ideal' subset of 95 NBI-N images selected by participating physicians for homogeneity to see ability of this network when tested in slightly more ideal cases of using near-focus NBI colonoscopes with good scale/centering on polyps. Examples of these ideal image cases are shown in Fig. 4.1.

4.4.1 Reconstruction Regularization Performance

As mentioned in Section 3.4.3, the idea of reconstructing the input as a method of regularization was used in CapsNet by Sabour et al. [105], and the theory behind this technique and the regularization effect it introduces is by mapping the capsule vectors back to the input data, this forces the network to pay attention to more relevant features about the input, which might not be as discriminative for the given task, yet still provide some useful information. We examine again the role of the reconstruction regularization branch of our new network structure. Given the deeper structure and

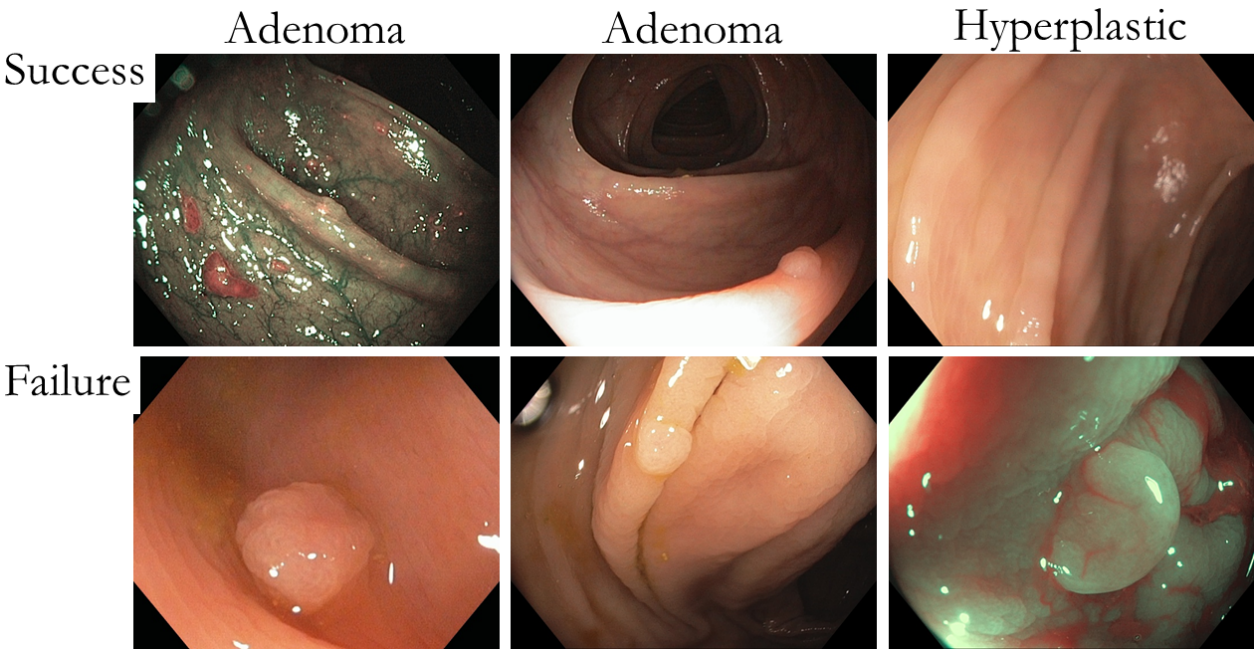


Figure 4.4: Qualitative evaluation for challenging examples: successful (first row) and failure (second row) cases are shown.

CAP operation introduced, it is important to examine what role reconstruction still plays in the network's performance. As evident by the results shown in Table 4.4 the reconstruction still plays an important role across all experiments, averaging an 8% decrease.

Table 4.4: Examining the effect on performance of reconstruction regularization in the diagnosis capsules framework. Table entries are accuracy percentages for experiments 1 – 3, namely Hyperplastic vs Adenoma (HP vs Ad), Hyperplastic vs Adenoma and Serrated (HP vs Ad & Ser), and Hyperplastic vs Serrated (HP vs Ser).

Method	HP vs Ad	HP vs Ad & Ser	HP vs Ser
No Recon	56%	50%	55%
With Recon	**66**%	**61**%	**58**%

4.4.2 Examining Dynamic Routing Iterations

Since the dynamic routing algorithm chosen for this study is an iterative process, we can investigate the optimal number of times to run the routing algorithm per forward pass of the network. In the original work by Sabour et al. [105], they found three iterations to provide the optimal results. We saw a consistent result for our *SegCaps* network architecture in Section 3.4.4. As seen in Table 4.5, the number of routing iterations has a similar effect on *D-Caps* predictive performance, where we find again that three iterations is optimal over a set of different numbers of iterations studied. The stability of this parameter is encouraging when applying the method to new application areas, leaving one less hyperparameter to tune per layer.

Table 4.5: Examining the effect on accuracy (%) of different number of routing iterations within *D-Caps* on Mayo Polyp dataset for Hyperplastic vs Adenoma.

Modality	Routings-2	Routings-3	Routings-4	Routings-5
All Images	54.73	**63.66**	53.37	51.85
All Polyps	50.61	**65.53**	45.97	50.86
NBI	51.18	**56.69**	50.66	50.92
NBIF	49.04	53.37	**59.13**	50.96
NBIN	47.14	**60.95**	59.52	53.33
WL	56.19	**68.81**	47.68	51.80
WLF	60.09	**72.48**	49.08	59.17
WLN	55.39	**67.65**	46.57	55.39
Near	48.20	**67.57**	50.90	54.50
Far	54.02	**69.64**	50.45	51.79

4.4.3 Testing on a More Ideal Subset

Given the considerable difficulty of the problem of diagnosing colorectal polyps in this extremely challenging Mayo Polyp dataset, where some performance was near random chance, we conducted an additional experiment in slightly more ideal conditions. An 'ideal' subset of 95 NBI-N images

was selected by participating physicians for homogeneity. These images are still from the Mayo Polyp dataset and contain many of the same challenges, but the physicans selected these images under the possible clinical workflow situation of a polyp being noted by the colonoscopist and a centered and larger scale close snapshot was taken. Examples of these images are shown in Fig. 4.1. In these more ideal cases we found that *D-Caps* achieved a significantly higher average accuracy of 82. This significantly higher accuracy gives strong encouragement to the future direction of a complete detection and diagnosis system which can first roughly localize polyps before attempting diagnosis.

4.5 Discussions & Conclusion on Capsule-Based Diagnosis

We designed a capsule-based network for the task of diagnosis in the field of endoscopy. In order to classify real-world imaging data much larger in size than those in MNIST or CIFAR, we introduced the concept of *capsule-average pooling*. Combining this with the memory saving advances made in Chapter 3, our proposed architecture *D-Caps* is able to diagnoses colorectal polyps from colonoscopy images. Given our preliminary evidence that capsule networks can better generalize to unseen poses, converges faster in training, and contains far fewer parameters than state-of-the-art CNNs, we hypothesized that *D-Caps* should be able to better handle the relatively limited training data and high intra-class variation present in the Mayo Polyp dataset. We conducted a set of thorough experiments to validate our hypothesis, stratified across all polyp categories, imaging devices and modalities, and focus modes available. Our results show *D-Caps* can outperform the leading state-of-the-art CNN-based method by as much as 43% in the most difficult settings. Ablation studies show that *D-Caps* can achieve performance beginning to approach clinical levels when given a slightly more ideal set of candidates to diagnose. Given this, future research directions into capsule-based object detection methods for colorectal polyps to give a more localized target for

diagnosis can likely prove significantly beneficial.

CHAPTER 5: ENCODING CAPSULES FOR EXPLAINABLE PREDICTIONS

Related Publications and Patents:

- Rodney LaLonde, Drew Torigian, and Ulas Bagci. "Encoding Visual Attributes in Capsules for Explainable Medical Diagnoses." *MICCAI 2020, International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2020. [74]
- Rodney LaLonde and Ulas Bagci. "Capsules for Image Analysis." *U.S. Patent Application 16/431,387; filed December 5, 2019*.

In this chapter, we further build off the work completed in Chapters 3 and 4 by tackling the important challenge of explainability in medical image diagnosis algorithms. The remainder of this chapter is organized into the following sections: Section 5.1 – A brief overview explainability in deep learning and lung cancer diagnosis; Section 5.2 – The *X-Caps* framework described in detail, including the modified dynamic routing algorithm and the multi-task framework for encoding visual-attribute information; Section 5.3 – Experiments conducted, our implementation settings, and their results; Section 5.4 – Ablation studies on the various components of our algorithms to determine the contribution of each aspect of our proposed method to the final results; and Section 5.5 – Discussions and concluding remarks.

5.1 An Overview of Explainability in Deep Learning & Lung Cancer Diagnosis

In machine learning, predictive performance typically comes at the cost of *interpretability* [11, 40, 68, 104]. While deep learning (DL) has played a major role in a wide array of fields, there exist several high-risk domains which have yet to be comparably impacted: military, security, transportation, finance, legal, and healthcare among others [10, 76, 95]. Deep neural networks are often called black-boxes due to their difficult-to-interpret decisions. This is characteristic of a deeper trend in machine learning, where predictive performance typically comes at the cost of *interpretability* [11, 40, 68, 104]. Although deep learning (DL) has played a major role in a wide array of fields, there exist several which have yet to be comparably impacted: military, security, transportation, finance, legal, and healthcare among others [10, 76, 95]. At its core, DL owes its success to the joining of two essential tasks, *feature extraction* and *feature classification*, learned in a joint manner, usually through a form of backpropagation. This was a step away from feature engineering, where experts would hand-craft the most import set of discernible characteristics for a task, while the classification of these features typically employed some form of machine learning. Although this direction has dramatically improved the predictive performance on a diverse range of tasks, it has also come at a great cost, the sacrifice of human-level *explainability*. As features becomes less *interpretable*, and the functions learned more complex, model predictions become more difficult to explain and the generalization ability of trained networks is less well understood. Using hand-crafted features, human-experts could know more precisely under what circumstances their algorithms would fail; this is no longer the case for most DL algorithms, where the generalization ability of trained networks is less well understood. Several works have began to press towards this goal of explainable DL, as explored in Section 2.3, but the problem remains largely unsolved.

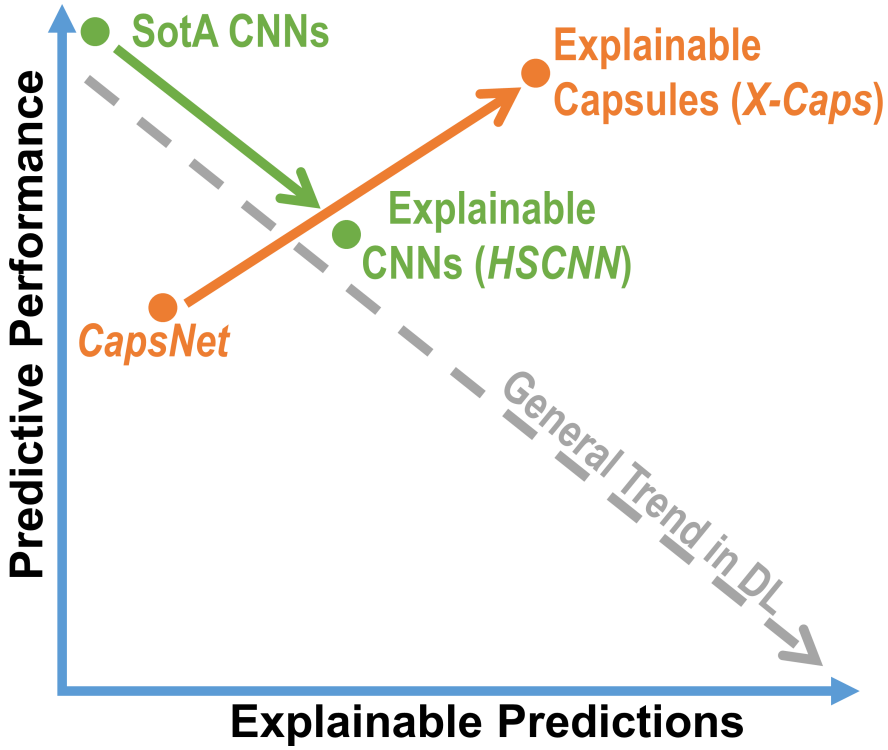


Figure 5.1: A symbolic plot showing the general trade-off between explainability and predictive performance in deep learning (DL) [11, 40, 68, 104]. Our proposed *X-Caps* rebuts the trend of decreasing performance from state-of-the-art (SotA) as explainability increases and shows it is possible to create more explainable models *and* increase predictive performance with capsule networks.

5.1.1 Interpretable vs. Explainable & Why Capsule Networks?

There has been a recent push in the community to move away from the *post-hoc interpretations* of deep models and instead create explainable models from the outset [104, 107]. Since the terms *interpretable* and *explainable* are often used interchangeably, we want to be explicit about our definitions for the purposes of this study. An *explainable* model is one which provides explanations for its predictions *at the human level* for a *specific task*. An *interpretable* model is one for which some conclusions can be drawn about the internals/predictions of the model; however, they are

not explicitly provided by the model and are typically at a lower level. For example, in image classification, when a deep model predicts an image to be of a cat, saliency/gradient or other methods can attempt to *interpret* the model/prediction. However, the model is not *explaining* why the object in the image is a cat in the same way as a human. Humans classify objects based on a taxonomy of characteristics/attributes (*e.g.* cat equals four legs, paws, whiskers, fur, etc.). If our goal is to create *explainable* models, we should design models which explain their decisions using a similar set of "attributes" to humans, instead of relying on class activation maps.

Why capsule networks? As stated in the previous chapters, capsule networks differ from convolutional neural networks (CNNs) by replacing the scalar feature maps with vectorized representations, responsible for encoding information (*e.g.* pose, scale, color) about each feature. These vectors are then used in a dynamic routing algorithm which seeks to maximize the agreement between lower-level predictions for the instantiation parameters (i.e. capsule vectors) of higher-level features. In their introductory work, a capsule network (*CapsNet*) was shown to produce promising results on the MNIST data set; but more importantly, was able to encode high-level visually-interpretable features of digits (*e.g.* stroke thickness, skew, localized-parts) within the dimensions of its capsule vectors [105]. In this way, capsules provide a good candidate for not only representing attributes with vectors, but being able to combine those attributes in meaningful ways to make decisions about the object being represented.

5.1.2 Lung Cancer: A High-Risk Application Needing Explainability

Lung cancer is the far-leading cause of cancer-related death in both men and women [36]. The National Lung Screening Trial showed that screening patients with low-dose computed tomography (CT) has reduced lung cancer mortality by 20% [117, 125]. However, only 16% of lung cancer cases are diagnosed at an early stage [51]. DL approaches such as 2D and 3D CNNs have been proposed

to alleviate these challenges. Noticeably, some have achieved highly successful diagnosis results, comparable to or even better than expert level diagnosis [54, 131]. Nevertheless, the black-box nature of these previous studies has contributed to these methods not making their way into clinical routines. The purpose of this study is to fill this important research gap by creating explainable medical diagnoses through learning visually-interpretable features from medical images with new DL models, specifically a novel capsule network architecture.

In diagnosing the malignancy of lung nodules, similar to describing why an image of a cat is catlike, radiologists explain their predictions through the language of high-level visual attributes (i.e., radiographical interpretations): subtlety (sub), sphericity (sph), margin (mar), lobulation (lob), spiculation (spi), and texture (tex), shown in Fig. 5.2, which are known to be predictive (with inherent uncertainty) of malignancy [45]. To create a DL model with this same level of radiographical interpretation, we propose a novel multi-task capsule architecture, called ***X-Caps***, for learning visually-interpretable feature representations within capsule vectors, then predicting malignancy based solely on these interpretable features. By supervising different capsules to embed specific visually-interpretable features, multiple visual attributes are learned simultaneously, with their weights being updated by both the radiologists visual interpretation scores as well as their contribution to the final malignancy score, regularized by the segmentation reconstruction error. Since these attributes are not mutually-exclusive, we introduce a new routing sigmoid function to independently route child capsules to parents. Further, to provide radiologists with an estimate of model confidence, we train our network on a distribution of expert labels, modeling inter-observer agreement and punishing over/under confidence during training, supervised by human-experts' agreement.

In this study, we show even a relatively simple 2D capsule network, *X-Caps*, can better capture high-level visual attribute information than the state-of-the-art deep dual-path dense 3D convolutional neural network (CNN) while also improving diagnostic accuracy, approaching that of even some

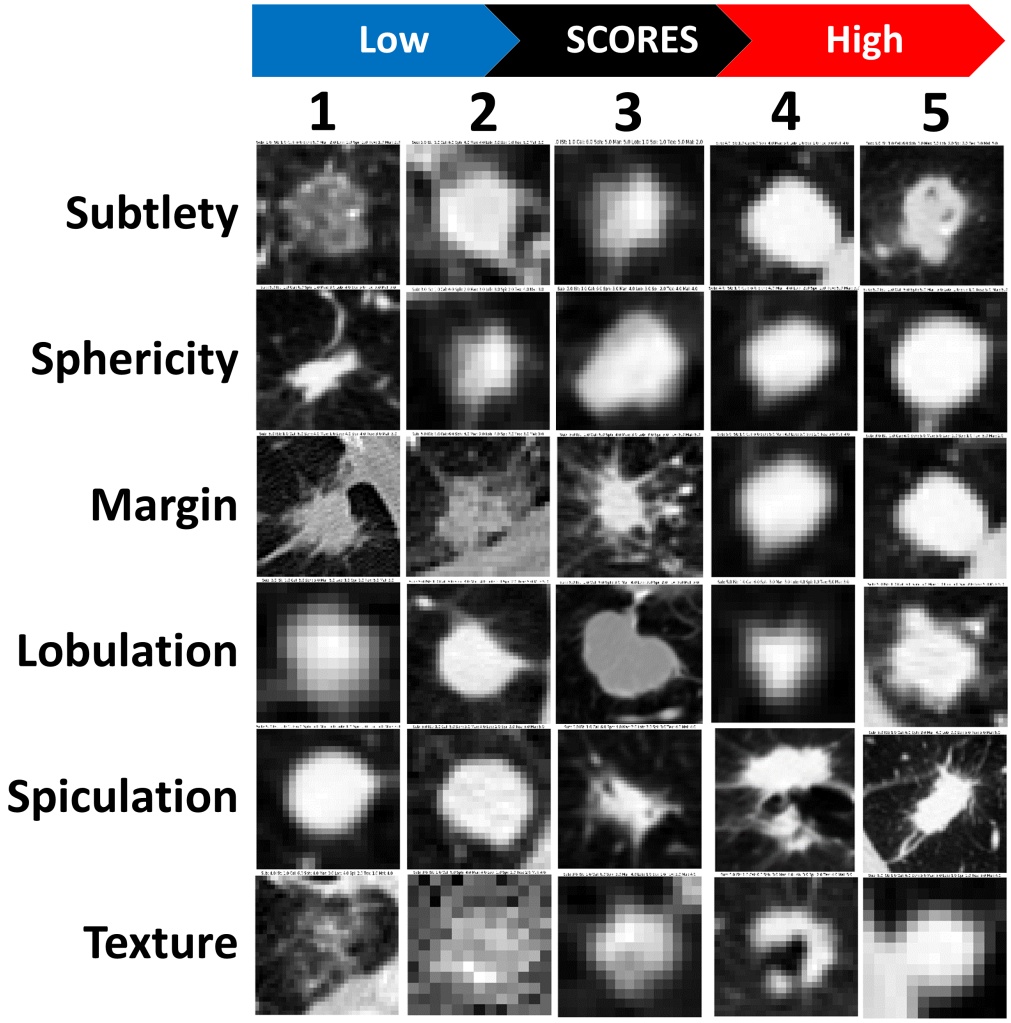


Figure 5.2: Lung nodules with high-level visual attribute scores as determined by expert radiologists. Scores were given from 1 – 5 for six different visual attributes related to diagnosing lung cancer.

black-box methods (*e.g.,* [108, 109]). Although we believe the proposed methods described are generic and can be applied to any classification problem in computer vision with visual attributes to be modeled, we choose to focus on a high-risk application area where explainability is a critical lynch-pin holding back the adoption of DL in routine use: lung cancer diagnosis.

Overall, the contributions of this study are summarized as:

1. The first study to directly encode high-level visual attributes within the vectors of a capsule network to perform explainable image-based diagnosis *at the radiologist-level*.

2. Create a novel modification the dynamic routing algorithm to independently route information from child capsules to parents when parent capsules are not mutually-exclusive.

3. Provide a meaningful confidence metric with our predictions at test by learning directly from expert label distributions to punish network over/under confidence. Visual attribute predictions are verified at test via the reconstruction branch of the network.

4. Demonstrate a simple 2D capsule network (*X-Caps*) trained from scratch outperforming a state-of-the-art deep pre-trained dual-path 3D dense CNN at capturing visually-interpretable high-level attributes and malignancy prediction, while providing malignancy prediction scores approaching that of non-explainable 3D CNNs.

5.2 Explaining Predictions by Encoding Attributes in Capsules

The goal of our proposed method is to model visual attributes using capsule neural networks for the important application domain of high-risk predictions. We apply our algorithm to CT lung data in order to provide the same explanations as radiologists for predicting malignancy, while simultaneously performing malignancy prediction and nodule segmentation/reconstruction. The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [6], described in more detail in Section 5.3, contains a collection of lung nodules with scores ranging from 1 – 5 across a set of visual attributes, indicating their relative appearance, and malignancy, as scored by up to four radiologists. These characteristics and scores are shown in Figure 5.2.

Our approach, referred to as *explainable capsules*, or *X-Caps*, was designed to remain as similar as possible to our control network, *CapsNet*, while allowing us to have more control over the

visually-interpretable features learned. *CapsNet* already showed great promise when trained on the MNIST data set for its ability to model high-level visually-interpretable features. With this study, we examine the ability of capsules to model *specific* visual attributes within their vectors, rather that simply hoping these are learned successfully in the more challenging lung nodule data. As shown in Figure 5.3, *X-Caps* shares a similar overall structure as *CapsNet*, with the major differences being the addition of the supervised labels for each of the *X-Caps* vectors, the fully-connected layer for malignancy prediction, the reconstruction regularization also performing segmentation, and the modifications to the dynamic routing algorithm.

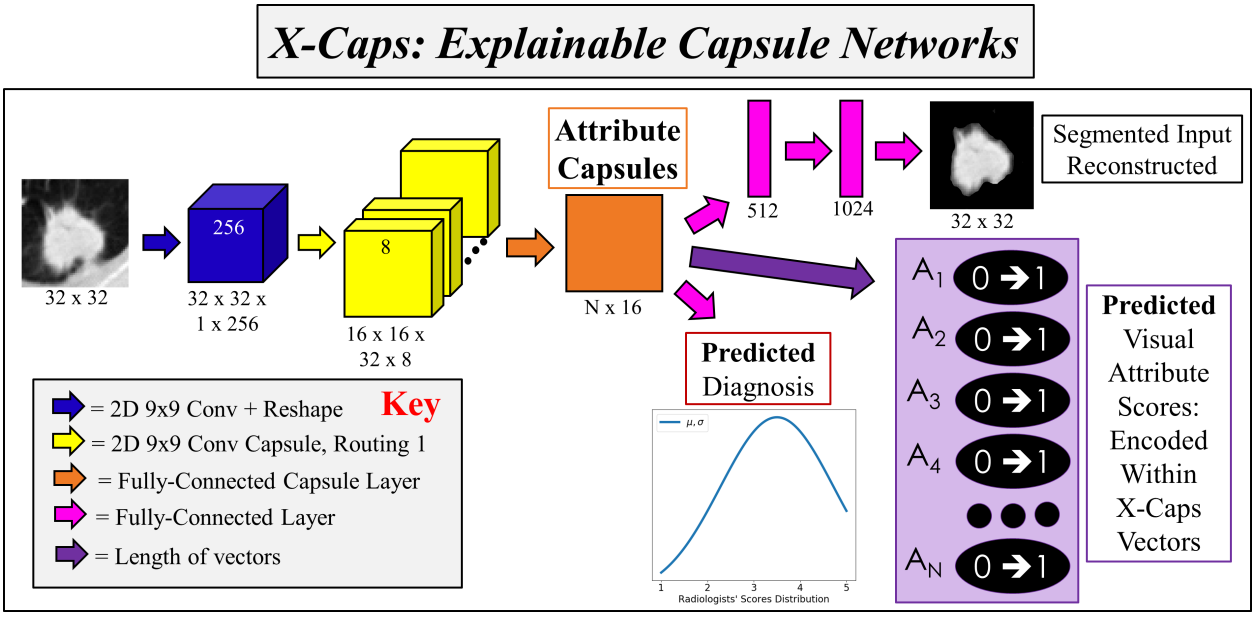


Figure 5.3: *X-Caps*: Explainable Capsule Networks. The proposed network (1) predicts *N* high-level visual attributes of the nodule, (2) segments the nodule and reconstruct the input image, and (3) diagnoses the nodule on a scale of 1 to 5 based on the visually-interpretable high-level features encoded in the X-Caps capsule vectors. The malignancy diagnosis branch is attempting to model the distribution of radiologists' scores in both mean and variance.

5.2.1 Building an Explainable Capsule Network

The first layer of our proposed network is a 2D convolutional layer which extracts the lowest-level features. Next, we form our primary capsules of 32 capsule types with $8D$ vector capsules. The primary capsules can be seen as either a convolution capsule layer with a single routing iteration or as grouping the feature maps of a convolutional layer and performing the non-linear squashing function from [105]. Following this, we form our attribute capsules using a fully-connected capsule layer whose output is N $16D$ capsule types, one for each of the visual-attributes we want to predict. Unlike *CapsNet* where each of the parent capsules were dependant on one another (e.g. if the prediction is the digit 5 it cannot also be a 3), our parent capsules are not mutually-exclusive of each other (i.e. a nodule can score high or low in each of the attribute categories). For this reason, we needed to modify the dynamic routing algorithm presented in *CapsNet* to accommodate this significant difference. The key change is the "routing softmax" employed by *CapsNet* forces the contributions of each child to send their information to parents in a manner which sums to one, which in practice effectively makes them "choose" a parent to send their information to. However, when computing prediction vectors for independent parents, we want a child to be able to contribute to all parent capsules for attributes which are present in the given input. With that motivation, the specific algorithm, which we call "routing sigmoid", is computed as

$$r_{i,j} = \frac{\exp(b_{i,j})}{\exp(b_{i,j}) + 1}, \tag{5.1}$$

where $r_{i,j}$ are the routing coefficients determined by the dynamic routing algorithm for child capsule i to parent capsule j and the initial logits, $b_{i,j}$ are the prior probabilities that the prediction vector for capsule i should be routed to parent capsule j. Note the prior probabilities are initially set to 1 rather than 0 as in *CapsNet*, otherwise no routing could take place. The rest of the dynamic routing procedure follows the same as in [105].

5.2.2 *Predicting Malignancy From Visually-Interpretable Encoded Capsules*

In order to predict malignancy scores, we attach a fully-connected layer to our X-Caps attribute prediction vectors with output size equal to the range of scores. We wish to emphasize here, our final malignancy prediction is coming solely from the vectors whose magnitudes represent *visually-interpretable* feature scores. Every malignancy prediction score has a set of weights connected to the high-level attribute capsule vectors, and the activation from each tells us the exact contribution of the given visual attribute to the final malignancy prediction for that nodule. Unlike previous studies which look at the importance of these attributes on a global level, our method looks at the importance of each visual attribute in relation to each specific nodule being diagnosed. To verify the correctness of our attribute modeling, we reconstruct the nodules while varying the dimensions of the capsule vectors to ensure the desired visual attributes are being modeled. At test, these reconstructions give confidence that the network is properly capturing the attributes, and thus the scores can be trusted. Confidence in the malignancy prediction score, in addition to coming solely from these trusted attributes, is provided via an uncertainty modeling approach.

Previous works in lung nodule classification follow the same strategy of averaging radiologists' scores for visual attributes and malignancy and then either attempt to regress this average or performing binary classification of the average as below or above 3. While such approaches make training simpler, they throw away valuable information about the agreement or disagreement among radiologists. To better model the uncertainty inherently present in the labels due to inter-observer variation, we propose a different approach: we attempt to predict the *distribution* of radiologists' scores. Specifically, for a given nodule where we have at minimum three radiologists' score values for each attribute and for malignancy prediction, we compute the mean and variance of those values and fit a Gaussian function to them, which is in turn used as the ground-truth for our classification vector. Nodules with strong inter-observer agreement produce a sharp peak, in which case wrong

or unsure (*i.e.,* low confidence score) predictions are severely punished. Likewise, for low inter-observer agreement nodules, we expect our network to output a more spread distribution and it will be punished for strongly predicting a single class label. This proposed approach allows us to model the uncertainty present in radiologists' labels in a way that no previous study has and provide a meaningful confidence metric at test time to radiologists.

5.2.3 Multi-Task Capsule Loss & Regularization

As in *CapsNet*, we also perform reconstruction of the input as a form of regularization. However, we extend the idea of regularization to perform a pseudo-segmentation, similar in nature to the reconstruction used by [71, 75]. Whereas in true segmentation, the goal is to output a binary mask of pixels which belong to the nodule region, in our formulation we attempt to reconstruct only the pixels which belong to the nodule region, while the rest are mapped to zero. More specifically, we formulate this loss as

$$\mathscr{L}_r = \frac{\gamma}{H \times W} \sum_{x}^{W} \sum_{y}^{H} \|R^{x,y} - O_r^{x,y}\|, \text{ with} \tag{5.2}$$

$$R^{x,y} = I^{x,y} \times S^{x,y} \mid S^{x,y} \in \{0,1\}, \tag{5.3}$$

where $\mathscr{L}_r$ is the supervised loss for the reconstruction regularization, γ is a weighting coefficient for the reconstruction loss, $R^{x,y}$ is the reconstruction target pixel, $S^{x,y}$ is the ground-truth segmentation mask value, and $O_r^{x,y}$ is the output of the reconstruction network, at pixel location (x,y), respectively, and H and W are the height and width, respectively, of the input image. This adds another task to our multi-task approach and an additional supervisory signal which can help our network distinguish visual characteristics from background noise. The malignancy prediction score, as well as each of the visual attribute scores also provide a supervisory signal in the form of

$$\mathscr{L}_a = \sum_{n}^{N} \alpha^n \|A^n - O_a^n\|, \text{ and} \tag{5.4}$$

$$\mathscr{L}_m = \beta \sum_{x \in X} \varepsilon(\boldsymbol{O}_m) \log \left(\frac{\varepsilon(\boldsymbol{O}_m)}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} \right), \tag{5.5}$$

where $\mathscr{L}_a$ is the combined loss for the visual attributes, A^n is the average of the attribute scores given by at minimum three radiologists for attribute n, N is the total number of attributes, α^n is the weighting coefficient placed on the n^{th} attribute, O_a^n is the network prediction for the score of the n^{th} attribute, $\mathscr{L}_m$ is a KL divergence loss for the malignancy score, β is the weighting coefficient for the malignancy score, μ and σ are the mean and variance of radiologists' scores, and $\varepsilon = \exp(O_m^i)/\sum_{j=1}^{N} \exp(O_m^j)$ is the softmax over the network malignancy prediction vector $\boldsymbol{O}_m = \{O_m^1, ..., O_m^N\}$. In this way, the overall loss for *X-Caps* is simply $\mathscr{L} = \mathscr{L}_m + \mathscr{L}_a + \mathscr{L}_r$. For simplicity, the values of each α^n and β are set to 1, and γ is set to $0.005 \times 32 \times 32 = 0.512$.[1].

5.3 Experiments in Explainable Lung Cancer Diagnosis & Results

For our experiments, we used publicly available LIDC-IDRI data set [6]. The LIDC-IDRI includes 1018 volumetric CT scans, where each CT scan was interpreted by at most four radiologists by the LIDC-IDRI project team. Lung nodules were given scores by participating radiologists for each of six visual attributes, shown in Fig. 5.2, and malignancy ranging from 1 to 5. For simplicity, and including malignancy indecision among radiologists, we excluded lung nodules from the consideration when their mean visual score was exactly 3. This left 1149 lung nodules to be evaluated (646 benign and 503 malignant). Table 5.1 shows the summary of visual score distribution of lung nodules evaluated by at least three radiologists.

Five-fold stratified cross-validation was performed to split the nodules into training and testing sets, with 10% of each training set set aside for validation and early stopping. All models were trained

[1]Further tuning of these parameters could potentially lead to superior results but we did not have the computational resources to perform such an analysis for this study.

Table 5.1: Numbers within the table represent individual radiologists' scores. At the nodule level, there were 1149 nodules after removing those with less than three radiologists and those with mean score 3: 646 benign (< 3.0) and 503 malignant (> 3.0) nodule were used for training and testing in cross-validation.

	Visual Attribute Scores				
Attributes	**1**	**2**	**3**	**4**	**5**
subtlety	124	274	827	1160	1817
sphericity	10	322	1294	1411	1165
margin	174	303	512	1362	1851
lobulation	2394	924	475	281	128
spiculation	2714	789	336	174	189
texture	207	76	188	485	3246
malignancy	676	872	1397	658	599

with a batch size of 16 using Adam with an initial learning rate of 0.02 reduced by a factor of 0.1 after validation loss plateau. All code is implemented in TensorFlow and has been made publicly available. Consistent with the literature, predictions were considered correct if within ± 1 of the radiologists' classification [54, 55].

The experimental results summarized in Table 5.2 illustrate the prediction of visual attributes with the proposed *X-Caps* in comparison with an adapted version of *CapsNet*, a deep dual-path dense 3D explainable CNN (*HSCNN* [107]), and two state-of-the-art non-explainable methods which do not have extra post-processing or learning strategies. Compared methods results are from the original reported works. To the best of our knowledge, *HSCNN* is the only other work in the literature which presents attribute-level predictions pursuant to creating explainable models through the modeling of high-level visual attributes for lung cancer diagnosis.

Our results show that a simple 2D capsule network has the ability to model visual attributes far better than *HSCNN* while also achieving better malignancy prediction. Further, we wish to emphasize the significance of *X-Caps* providing increased predictive performance *and* explainability over *CapsNet*. This goes against the assumed trend in DL, illustrated with a symbolic plot in

Table 5.2: Prediction accuracy of visual attribute learning with capsule networks. Dashes (-) represent values which the given method could not produce. *X-Caps* outperforms the state-of-the-art explainable method (*HSCNN*) at attribute modeling (the main goal of both studies), while also producing higher malignancy prediction scores, approaching state-of-the-art non-explainable methods performance.

	Attribute Prediction Accuracy %						**Malig-**
	sub	sph	mar	lob	spi	tex	**nancy**
Non-Explainable Methods							
3D Multi-Scale + RF [108]	-	-	-	-	-	-	86.84
3D Multi-Crop [109]	-	-	-	-	-	-	87.14
CapsNet [105]	-	-	-	-	-	-	77.04
Explainable Methods							
3D Dual-Path *HSCNN* [107]	71.9	55.2	72.5	-	-	83.4	84.20
Proposed *X-Caps*	**90.39**	**85.44**	**84.14**	**70.69**	**75.23**	**93.10**	**86.39**

Figure 5.1, that explainability comes at the cost of predictive performance, a trend we observe with *HSCNN* being outperformed by less powerful (*i.e.* not dense or dual-path) but non-explainable 3D CNNs [108, 109]. We can observe this trend in the example of [109] achieving 87% accuracy with a 3D CNN and [107] achieving 84% accuracy with an objectively more powerful dense dual-path 3D CNN but which also provides explainable predictions in the form of high-level visual attributes. While *X-Caps* slightly under-performs the best non-explainable models, it is reasonable to suspect that future research into deeper and more powerful 3D capsule networks with tricks like residual or dense connections, normalization, and other novelties which were introduced to CNNs over the last several years, would allows explainable capsules to surpass these methods; we hope this study will promote such future works.

5.4 Ablation Studies for the Components of *X-Caps*

To analyze the impact of each component of our proposed approach, we performed ablation studies for: (1) learning the distribution of radiologists' scores rather than attempting to regress the mean

Table 5.3: Ablation studies for malignancy prediction accuracy: (1) regressing the mean score instead of predicting the distribution, (2) no reconstruction regularization, (3) using *CapsNet*'s "routing softmax" instead of the proposed "routing sigmoid", and (4) the proposed approach.

Mean Score	No Recon.	Routing Softmax	Proposed Method
83.09%	80.30%	80.69%	**86.39%**

value of these scores, (2) removing the reconstruction regularization from the network, and (3) performing our proposed "routing sigmoid" over the original "routing softmax" proposed in [105].

The results of each of these ablations is shown in Table 5.3 and we can see removing each component had a significant negative impact on our malignancy prediction results. This shows retaining the agreement/disagreement information among radiologists proved significantly useful, the reconstruction played a role in improving the network performance, and our proposed modifications to the dynamic routing algorithm were necessary for passing information from children to parents when the parent capsule types are independent.

As limitations of our work, we did not tune the weight balancing terms between the different tasks and further investigation could lead to superior performance. Also, we found capsule networks can be somewhat fragile; often random initializations failed to converge to good performance. However, this might be due to the small/shallow network size and its relation to the Lottery Ticket Hypothesis [37] rather than anything specific to capsules. Lastly, although we defined the loss in Section 5.2.3 as mean squared error, we also experimented with cross-entropy, margin, and Kullback-Leibler divergence loss functions. From our empirical analysis, these loss functions all performed comparably with each other, although a more systematic investigation would be needed to draw any firm conclusions.

5.5 Discussions & Conclusion on Explainable Deep Learning With Capsules

Available studies for explaining DL models, typically focus on *post hoc* interpretations of trained networks, rather than attempting to build-in explainability. This is the first study for directly learning an interpretable feature space by encoding high-level visual attributes within the vectors of a capsule network to perform explainable image-based diagnosis. We approximate visually-interpretable attributes through individual capsule types, then predict malignancy scores directly based only on these high-level attribute capsule vectors, in order to provide malignancy predictions with explanations *at the human-level*, in the same language used by radiologists. Our proposed multi-task explainable capsule network, *X-Caps*, successfully approximated visual attribute scores better than the previous state-of-the-art explainable diagnosis system, while also achieving higher diagnostic accuracy. We hope our work can provide radiologists with malignancy predictions which are explained via the same high-level visual attributes they currently use, while also providing a meaningful confidence metric to advise when the results can be more trusted, thus allowing radiologists to quickly interpret and verify our predictions. Lastly, although we selected lung cancer diagnosis for testing our method, we believe our approach should be generally applicable to any image-based classification task where high-level attribute information is available to provide explanations about the final prediction.

CHAPTER 6: CONCLUSION & FUTURE DIRECTIONS

The body of work presented in this dissertation constitutes significant algorithmic advances to the application of capsule networks in a variety of real-world imaging data domains, and in particular, biomedical image computer-aided diagnosis. First, we provide a brief discussion on each main chapter of this document, summarizing the novelties introduced and the empirical evidence gained through our experimental results. Following this, we discuss the main research areas and topics which would be of significant interest given the insights garnered from the summarized body of work.

6.1 Final Conclusions

In Chapter 3 we introduced the first ever capsule-based segmentation network in the literature, *SegCaps*, while producing several important advancements, including a novel locally-constrained dynamic routing algorithm, transformation matrix sharing, the concept of a "deconvolutional" capsule, extension of the reconstruction regularization to segmentation, and a new encoder-decoder capsule network structure. We validated the effectiveness and efficiency of our proposed method against several state-of-the-art CNN-based methods in the largest ever study in pathological lung segmentation, and the only showing results on pre-clinical subjects utilizing deep learning methods. *SegCaps* consistently outperforms all other compared state-of-the-art approaches in terms of the commonly measured metrics, Dice and HD. Additionally, *SegCaps* achieves this while only using a fraction of the total parameters of these much larger networks. Our results in the main body of experiments, as well as the additional experiments conducted on other objects in other data modalities, give compelling evidence for the advantages of a capsule-based segmentation method over CNN-based methodologies.

In Chapter 4, we design a capsule-based diagnosis network, *D-Caps*, by introducing a novel capsule-average pooling technique, which we show can better handle the relatively limited training data and high intra-class variation present in colorectal cancer diagnosis. Combining this with the memory saving advances made in Chapter 3, our proposed architecture is able to classify real-world imaging data much larger in size than those in MNIST or CIFAR, diagnosing colorectal polyps from colonoscopy images. Given our preliminary evidence that capsule networks can better generalize to unseen poses, converges faster in training, and contains far fewer parameters than state-of-the-art CNNs, we hypothesized that *D-Caps* should be able to better handle the relatively limited training data and high intra-class variation present in the Mayo Polyp dataset. We conducted a set of thorough experiments to validate our hypothesis, stratified across all polyp categories, imaging devices and modalities, and focus modes available. Our results show *D-Caps* can outperform the leading state-of-the-art CNN-based method by as much as 43% in the most difficult settings. Ablation studies show that *D-Caps* can achieve performance beginning to approach clinical levels when given a slightly more ideal set of candidates to diagnose.

In Chapter 5, we design an explainable capsule network, *X-Caps*, which encodes high-level visual object attributes within the vectors of its capsules, then forms explanations for its predictions using the same high-level language used by human-experts in a multi-task learning framework. Utilizing a novel modification to the dynamic routing algorithm which independently routes information from child capsules to parents, we train our network directly on the distribution of expert labels, modeling inter-observer agreement, and thus providing a meaningful metric of model over/under confidence supervised by human-experts' agreement. We demonstrate a simple 2D capsule network trained from scratch can outperform a state-of-the-art deep pre-trained dense dual-path 3D CNN at capturing visually-interpretable high-level attributes and malignancy prediction, while providing malignancy prediction scores approaching that of non-explainable 3D CNNs.

Capsule networks show considerable promise for the future of deep learning-based applications and

we hope the contributions of this dissertation provide a solid foundation for the further advancement of capsule network-based approaches. The source code for all of the algorithms discussed have been made publicly available at https://github.com/lalonderodney.

6.2 Future Research Directions

As with most research, there are two main thrusts which can be pursued, technical advancements and novel applications. We first discuss the most crucial technical shortcomings of capsule networks and our recommendations to future researchers for possible directions of investigation. Following this we discuss the application domains which have been least impacted by capsule networks as a possible motivation for further exploration by future researchers.

From a technical point-of-view, the most important technical advancements within capsule networks need to be made in the dynamic routing mechanism. Dynamic routing was cited as a key contribution by Hinton to finally make capsule networks a reality; however, there is strong evidence that the current iterative-based routing mechanisms in the literature are significantly sub-optimal [91]. Some interesting work has been completed in this area, including by Kosiorek et al. [66] and Tsai et al. [118], but there is still much to be desired. We would encourage future researchers to push along the lines of representation learning/disentanglement [9, 17], where capsule networks share some very interesting parallels to concept-vector based methods [39, 62] and ideas such as concept whitening [18]. Not only would such approaches allow us to more intelligently route information through the capsule network, but would provide significant improvements into explainability of deep methods.

On the application side of research, the majority of previous investigations have focused on image-based classification [21, 49, 97, 122]. Some notable exceptions are the small body of work which has investigated capsule networks in action detection [24, 25, 84], point-cloud autoencoders [129],

adversarial detection [38, 96], similarity matching/image retrieval [52, 65], generative methods [57], and reinforcement learning [5]. One application which is noticeably missing from this list is object detection in images. There is a rather large segment of the computer vision community which focuses on methods for object detection, and it is somewhat surprising that no published research has emerged to solve the task utilizing a capsule-based network. We would encourage future researchers to attempt to create a capsule network for object detection to rival the big names in detection, such as R-CNN [41, 42, 101] and YOLO [98, 99, 100].

LIST OF REFERENCES

[1] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `http://tensorflow.org/`. Software available from tensorflow.org.

[2] P. Afshar, A. Mohammadi, and K. N. Plataniotis. Brain tumor type classification via capsule networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3129–3133, 2018. doi: 10.1109/ICIP.2018.8451379.

[3] M. A. Alcorn et al. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2019.

[4] American Cancer Society. Key statistics for colorectal cancer. `www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html`, 2018.

[5] P.-A. Andersen. Deep reinforcement learning using capsules in advanced game environments. *arXiv preprint arXiv:1801.09597*, 2018.

[6] S. G. Armato et al. The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics*, 38(2):915–931, 1 2011. ISSN 0094-2405. doi: 10.1118/1.3528204.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[8] U. Bagci, X. Chen, and J. K. Udupa. Hierarchical scale-based multiobject recognition of 3-d anatomical structures. *IEEE Transactions on Medical Imaging*, 31(3):777–789, 2012.

[9] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 3319–3327, July 2017. doi: 10.1109/CVPR.2017.354. URL doi.ieeecomputersociety.org/10.1109/CVPR.2017.354.

[10] J. Bloomberg. Don't Trust Artificial Intelligence? Time To Open The AI 'Black Box'. http://www.forbes.com/sites/jasonbloomberg/2018/09/16/dont-trust-artificial-intelligence-time-to-open-the-ai-black-box/#6ceaf3793b4a, 11.16.2018. Forbes Magazine.

[11] G. Bologna. A model for single and multiple knowledge based networks. *Artificial Intelligence in Medicine*, 28(2):141–163, 2003.

[12] H. Brenner, C. Stock, and M. Hoffmeister. Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. *BMJ*, 348, 2014. doi: 10.1136/bmj.g2467. URL https://www.bmj.com/content/348/bmj.g2467.

[13] M. Buty, Z. Xu, M. Gao, U. Bagci, A. Wu, and D. J. Mollura. Characterization of lung nodule malignancy using hybrid shape and appearance features. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 662–670. Springer, 2016.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[16] P.-J. Chen, M.-C. Lin, M.-J. Lai, J.-C. Lin, H. H.-S. Lu, and V. S. Tseng. Accurate classification of diminutive colorectal polyps using computer-aided analysis. *Gastroenterology*, 154(3):568 – 575, 2018. ISSN 0016-5085. doi: https://doi.org/10.1053/j.gastro.2017.10.010. URL http://www.sciencedirect.com/science/article/pii/S0016508517362510.

[17] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[18] Z. Chen, Y. Bei, and C. Rudin. Concept whitening for interpretable image recognition. *arXiv preprint arXiv:2002.01650*, 2020.

[19] F. Chollet et al. Keras. https://keras.io, 2015.

[20] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[21] F. Deng, S. Pu, X. Chen, Y. Shi, T. Yuan, and S. Pu. Hyperspectral image classification with capsule network using limited training samples. *Sensors*, 18(9):3153, 2018.

[22] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P.-A. Poletti, and H. Müller. Building a reference multimedia database for interstitial lung diseases. *Computerized Medical Imaging and Graphics*, 36(3):227 – 238, 2012. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2011.07.003. URL http://www.sciencedirect.com/science/article/pii/S0895611111001017.

[23] R. Dey, Z. Lu, and Y. Hong. Diagnostic classification of lung nodules using 3d neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 774–778. IEEE, 2018.

[24] K. Duarte, Y. Rawat, and M. Shah. Videocapsulenet: A simplified network for action detection. In *Advances in Neural Information Processing Systems*, pages 7610–7619, 2018.

[25] K. Duarte, Y. S. Rawat, and M. Shah. Capsulevos: Semi-supervised video object segmentation using capsule routing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8480–8489, 2019.

[26] C. J. R. et al. Narrow band imaging optical diagnosis of small colorectal polyps in routine clinical practice: the detect inspect characterise resect and discard 2 (discard 2) study. *Gut*, 66:887–895, 2017. ISSN 0017-5749. doi: 10.1136/gutjnl-2015-310584. URL `https://gut.bmj.com/content/66/5/887`.

[27] J. E. G. I. et al. Development and validation of the wasp classification system for optical diagnosis of adenomas, hyperplastic polyps and sessile serrated adenomas/polyps. *Gut*, 65(6):963–970, 2016. ISSN 0017-5749. doi: 10.1136/gutjnl-2014-308411. URL `https://gut.bmj.com/content/65/6/963`.

[28] M. B. W. et al. Accuracy of in vivo colorectal polyp discrimination by using dual-focus high-definition narrow-band imaging colonoscopy. *Gastrointestinal Endoscopy*, 80:1072 – 1087, 2014. ISSN 0016-5107. doi: https://doi.org/10.1016/j.gie.2014.05.305. URL `http://www.sciencedirect.com/science/article/pii/S0016510714017957`.

[29] M. F. B. et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut*, 2017. ISSN 0017-5749. doi: 10.1136/gutjnl-2017-314547. URL `https://gut.bmj.com/content/early/2017/11/09/gutjnl-2017-314547`.

[30] P. M. et al. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE Transactions on Medical Imaging*, 35(9):2051–2063, Sept 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2547947.

[31] R. Z. et al. Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain. *IEEE Journal of Biomedical and Health Informatics*, 21(1):41–47, Jan 2017. ISSN 2168-2194. doi: 10.1109/JBHI.2016.2635662.

[32] Y. K. et al. Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy. *Gastrointestinal Endoscopy*, 83(3):643 – 649, 2016. ISSN 0016-5107. doi: https://doi.org/10.1016/j.gie.2015.08.004. URL `http://www.sciencedirect.com/science/article/pii/S0016510715027388`.

[33] Y. S. et al. Narrow-band imaging (nbi) magnifying endoscopic classification of colorectal tumors proposed by the japan nbi expert team. *Digestive Endoscopy*, 28(5):526–533, 2016. doi: 10.1111/den.12644. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/den.12644`.

[34] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. `http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html`, 2012.

[35] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

[36] N. C. for Health Statistics (US et al. Health, united states, 2016: with chartbook on long-term trends in health, 2017.

[37] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[38] N. Frosst, S. Sabour, and G. Hinton. Darccc: Detecting adversaries by reconstruction from class conditional capsules. *arXiv preprint arXiv:1811.06969*, 2018.

[39] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pages 9273–9282, 2019.

[40] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[41] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[42] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[43] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[44] L. Grady. Random walks for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 28(11):1768–1783, 2006.

[45] M. Hancock and J. Magnan. Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms. *Journal of Medical Imaging*, 3(4):044504, 2016.

[46] A. P. Harrison, Z. Xu, K. George, L. Lu, R. M. Summers, and D. J. Mollura. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from ct images. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 621–629, Cham, 2017. Springer International Publishing. ISBN 978-3-319-66179-7.

[47] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[48] G. Hinton. What is wrong with convolutional neural nets? `https://techtv.mit.edu/collections/bcs/videos/30698-what-s-wrong-with-convolutional-nets`, 12.04.2014. MIT Brain & Cognitive Sciences, lecture notes, Fall Colloquium Series.

[49] G. E. Hinton, S. Sabour, and N. Frosst. Matrix capsules with EM routing. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=HJWLfGWRb`.

[50] J. Horowitz and T. Pavlidis. Picture segmentation by a direct split-and-merge procedure. In *Proc. of the 2nd Int. Joint Conf. on Pattern Recognition*, pages 424–433, 1974.

[51] N. Howlader, A. Noone, M. Krapcho, D. Miller, K. Bishop, S. Altekruse, C. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D. Lewis, H. Chen, E. Feuer, and K. Cronin. SEER Cancer Statistics Review, 1975-2013, National Cancer Institute. https://seer.cancer.gov/archive/csr/1975_2013/, 04.2018.

[52] W.-L. Hsiao and K. Grauman. Creating capsule wardrobes from fashion images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7161–7170, 2018.

[53] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2017.

[54] S. Hussein, K. Cao, Q. Song, and U. Bagci. Risk stratification of lung nodules using 3d

cnn-based multi-task learning. In *International Conference on Information Processing in Medical Imaging*, pages 249–260. Springer, 2017.

[55] S. Hussein, R. Gillies, K. Cao, Q. Song, and U. Bagci. Tumornet: Lung nodule characterization using multi-view convolutional neural network with gaussian process. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 1007–1010. IEEE, 2017.

[56] T. Iesmantas and R. Alzbutas. Convolutional capsule network for classification of breast cancer histology images. In *International Conference Image Analysis and Recognition*, pages 853–860. Springer, 2018.

[57] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan. Capsulegan: Generative adversarial capsule network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[58] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1175–1183. IEEE, 2017.

[59] A. Jiménez-Sánchez, S. Albarqouni, and D. Mateus. Capsule networks against medical imaging data challenges. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 150–160. Springer, 2018.

[60] P. Kandel, R. LaLonde, V. Ciofoaia, M. B. Wallace, and U. Bagci. Su1741 colorectal polyp diagnosis with contemporary artificial intelligence. *Gastrointestinal Endoscopy*, 89(6): AB403, 2019.

[61] R. A. Karwoski, B. Bartholmai, V. A. Zavaletta, D. Holmes, and R. A. Robb. Processing of ct images for analysis of diffuse lung disease in the lung tissue research consortium. In *Medical Imaging 2008: Physiology, Function, and Structure from Medical Images*, volume 6916, page 691614. International Society for Optics and Photonics, 2008.

[62] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279*, 2017.

[63] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne. Learning how to explain neural networks: Patternnet and patternattribution. In *International Conference on Learning Representations (ICLR)*, 2018.

[64] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[65] F. Kınlı, B. Özcan, and F. Kıraç. Fashion image retrieval with capsule networks. *arXiv preprint arXiv:1908.09943*, 2019.

[66] A. R. Kosiorek, S. Sabour, Y. W. Teh, and G. E. Hinton. Stacked capsule autoencoders. *arXiv preprint arXiv:1906.06818*, 2019.

[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[68] M. Kuhn and K. Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.

[69] D. Kumar, A. Wong, and G. W. Taylor. Explaining the unexplained: A class-enhanced attentive response (clear) approach to understanding deep neural networks. In *IEEE Computer Vision and Pattern Recognition (CVPR) Workshop*, 2017.

[70] H. Lakkaraju, E. Kamar, R. Caruana, and E. Horvitz. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *AAAI*, pages 2124–2132, 2017.

[71] R. LaLonde and U. Bagci. Capsules for object segmentation. *arXiv preprint arXiv:1804.04241*, 2018.

[72] R. LaLonde, D. Zhang, and M. Shah. Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information. In *Computer Vision and Pattern Recognition, 2018. CVPR 2018. IEEE Computer Society Conference on*, 2018.

[73] R. LaLonde, P. Kandel, C. Spampinato, M. B. Wallace, and U. Bagci. Diagnosing colorectal polyps in the wild with capsule networks. In *17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020.

[74] R. LaLonde, D. Torigian, and U. Bagci. Encoding visual attributes in capsules for explainable medical diagnoses. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Cham, 2020. Springer International Publishing.

[75] R. LaLonde, Z. Xu, S. Jain, and U. Bagci. Capsules for biomedical image segmentation. *arXiv preprint arXiv:2004.04736*, 2020.

[76] M. Lehnis. Can We Trust AI If We Don't Know How It Works? http://www.bbc.com/news/business-44466213, 15.06.2018. BBC News.

[77] X. Li, Y. Kao, W. Shen, X. Li, and G. Xie. Lung nodule malignancy prediction using multi-task convolutional neural network. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 1013424. International Society for Optics and Photonics, 2017.

[78] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for

high-resolution semantic segmentation. In *Computer Vision and Pattern Recognition, 2017. CVPR 2017. IEEE Computer Society Conference on*, 2017.

[79] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Computer Vision and Pattern Recognition, 2015. CVPR 2015. IEEE Computer Society Conference on*, pages 3431–3440, 2015.

[80] J. L. Long, N. Zhang, and T. Darrell. Do convnets learn correspondence? In *Advances in Neural Information Processing Systems*, pages 1601–1609, 2014.

[81] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[82] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[83] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[84] B. McIntosh, K. Duarte, Y. S. Rawat, and M. Shah. Multi-modal capsule routing for actor and action video segmentation conditioned on natural language queries. *arXiv preprint arXiv:1812.00303*, 2018.

[85] A. Mobiny and H. Van Nguyen. Fast capsnet for lung cancer screening. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 741–749. Springer, 2018.

[86] A. Mobiny, H. Lu, H. V. Nguyen, B. Roysam, and N. Varadarajan. Automated classification

of apoptosis in phase contrast microscopy using capsule network. *IEEE transactions on medical imaging*, 39(1):1–10, 2019.

[87] A. Mortazi, J. Burt, and U. Bagci. Multi-planar deep segmentation networks for cardiac substructures from mri and ct. *stat*, 1050:3, 2017.

[88] A. Mortazi, R. Karim, K. Rhode, J. Burt, and U. Bagci. Cardiacnet: Segmentation of left atrium and proximal pulmonary veins from mri using multi-view cnn. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 377–385. Springer, 2017.

[89] A. Nibali, Z. He, and D. Wollersheim. Pulmonary nodule classification with deep residual networks. *International journal of computer assisted radiology and surgery*, 12(10):1799–1808, 2017.

[90] W. Nie, Y. Zhang, and A. Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International Conference on Machine Learning*, pages 3806–3815, 2018.

[91] I. Paik, T. Kwak, and I. Kim. Capsule networks need an improved routing algorithm. *arXiv preprint arXiv:1907.13327*, 2019.

[92] A. Pal, A. Chaturvedi, U. Garain, A. Chandra, R. Chatterjee, and S. Senapati. Capsdemm: capsule network for detection of munro's microabscess in skin biopsy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 389–397. Springer, 2018.

[93] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Computer Vision and Pattern Recognition, 2017. CVPR 2017. IEEE Computer Society Conference on*, pages 4353–4361, 2017.

[94] D. L. Pham, C. Xu, and J. L. Prince. Current methods in medical image segmentation. *Annual review of biomedical engineering*, 2(1):315–337, 2000.

[95] V. Polonski. People Don't Trust AI–Here's How We Can Change That. http://www.scientificamerican.com/article/people-dont-trust-ai-heres-how-we-can-change-that/, 10.01.2018. Scientific American.

[96] Y. Qin, N. Frosst, S. Sabour, C. Raffel, G. Cottrell, and G. Hinton. Detecting and diagnosing adversarial images with class-conditional capsule reconstructions. *arXiv preprint arXiv:1907.02957*, 2019.

[97] J. Rajasegaran, V. Jayasundara, S. Jayasekara, H. Jayasekara, S. Seneviratne, and R. Rodrigo. Deepcaps: Going deeper with capsule networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[98] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[99] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[100] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[101] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[102] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical

image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[103] A. Rosenfeld and A. C. Kak. *Digital Picture Processing: Volume 1*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2 edition, 1982. ISBN 9780323139915.

[104] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[105] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866, 2017.

[106] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[107] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, and W. Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Systems with Applications*, 2019.

[108] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian. Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging*, pages 588–599. Springer, 2015.

[109] W. Shen, M. Zhou, F. Yang, D. Yu, D. Dong, C. Yang, Y. Zang, and J. Tian. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, 61:663–673, 2017.

[110] Y. Shen and M. Gao. Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization. In *International Workshop on Machine Learning in Medical Imaging*, pages 389–397. Springer, 2018.

[111] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[112] C. Sorensen. How U of T's 'godfather' of deep learning is reimagining AI. `https://www.utoronto.ca/news/how-u-t-s-godfather-deep-learning-reimagining-ai`, 11.02.2017. U of T News.

[113] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[114] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pages 3308–3318, 2017.

[115] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[116] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[117] N. L. S. T. R. Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

[118] Y.-H. H. Tsai, N. Srivastava, H. Goh, and R. Salakhutdinov. Capsules with inverted dot-product attention routing. *arXiv preprint arXiv:2002.04764*, 2020.

[119] J. K. Udupa and S. Samarasekera. Fuzzy connectedness and object definition: theory, algorithms, and applications in image segmentation. *Graphical models and image processing*, 58(3):246–261, 1996.

[120] J. Vanian. Eye on A.I. — Celebrating the Godfathers of Deep Learning. `https://fortune.com/2019/04/02/eye-on-ai-godfathers-deep-learning/`, 04.02.2019. Fortune.

[121] L. A. Vese and T. F. Chan. A multiphase level set framework for image segmentation using the mumford and shah model. *International journal of computer vision*, 50(3):271–293, 2002.

[122] C. Xiang, L. Zhang, Y. Tang, W. Zou, and C. Xu. Ms-capsnet: A novel multi-scale capsule network. *IEEE Signal Processing Letters*, 25(12):1850–1854, 2018.

[123] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proceedings of IEEE International Conference on Computer Vision*, 2015.

[124] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692, 2018.

[125] D. F. Yankelevitz and J. P. Smith. Understanding the core result of the national lung screening trial. *New England Journal of Medicine*, 368(15):1460–1461, 2013.

[126] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[127] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, and S.-C. Zhu. Interpreting cnn knowledge via an explanatory graph. In *AAAI*, 2018.

[128] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Computer Vision and Pattern Recognition, 2017. CVPR 2017. IEEE Computer Society Conference on*, pages 2881–2890, 2017.

[129] Y. Zhao, T. Birdal, H. Deng, and F. Tombari. 3d point capsule networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1009–1018, 2019.

[130] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. In *International Conference on Learning Representations (ICLR)*, 2015.

[131] W. Zhu, C. Liu, W. Fan, and X. Xie. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 673–681. IEEE, 2018.

[132] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *International Conference on Learning Representations (ICLR)*, 2017.